DOCUMENT RESUME

ED 123 678 CS 501 389

TITLE	Status Report on Speech Research: A Report on the Status and Progress of Studies on the Nature of Speech, Instrumentation for Its Investigation, and Practical Applications, January 1 - June 30, 1976.
INSTITUTION	Haskins Labs., New Haven, Conn.
REPORT NO	SR-45/46-(1976)
PUB DATE	76
NOTE	238p.
EDRS PRICE	MF-$0.83 HC-$12.71 Plus Postage.
DESCRIPTORS	Articulation (Speech); Beginning Reading; Conference Reports; *Educational Research; Higher Education; Language Development; *Language Skills; *Oral Communication; *Speech; Speech Skills
IDENTIFIERS	*Status Reports

ABSTRACT

This report, covering the period of January 1 to June 30, 1976, is one of a regular series on the status and progress of studies on the nature of speech, instrumentation for its investigation, and practical applications. The manuscripts and extended reports contained in this report include "Exploring the Relations between Reading and Speech," "On Interpreting the Error Pattern in Beginning Reading," "Comments on the Session: Perception and Production of Speech II; Conference on Origins and Evolution of Language and Speech," "Consonant Environment Specifies Vowel Identity," "What Information Enables a Listener to Map a Talker's Vowel Space?" "Identification of Dichotic Fusions," "Discrimination of Dichotic Fusions," "Coperception: Two Further Preliminary Studies," "'Posner's Paradigm' and Categorical Perception: A Negative Study," "Weak Syllables in a Primitive Reading-Machine Algorithm." "Control of Fundamental Frequency, Intensity, and Register of Phonation," "The Effect of Delayed Auditory Feedback on Phonation: An Electromyographic Study," "Some Aspects of Coarticulation," "The Function of Strap Muscles in Speech," and "Laryngeal Muscle Activity in Stuttering." (RB)

****
* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. Nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *
****

ED123678

U S DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

SR-45/46 (1976)

Status Report on

SPEECH RESEARCH

A Report on
the Status and Progress of Studies on
the Nature of Speech, Instrumentation
for its Investigation, and Practical
Applications

1 January - 30 June 1976

Haskins Laboratories
270 Crown Street
New Haven, Conn. 06510

Distribution of this document is unlimited.

(This document contains no information not freely available to the general public. Haskins Laboratories distributes it primarily for library use. Copies are available from the National Technical Information Service or the ERIC Document Reproduction Service. See the Appendix for order numbers of previous Status Reports.)

CS 501 389

ACKNOWLEDGMENTS

The research reported here was made possible in part by support from the following sources:

National Institute of Dental Research
Grant DE-01774

National Institute of Child Health and Human Development
Grant HD-01994

Assistant Chief Medical Director for Research and Development, Research Center for Prosthetics, Veterans Administration
Contract V101(134)P-342

Advanced Research Projects Agency, Information Processing Technology Office, under contract with the Office of Naval Research, Information Systems Branch
Contract N00014-76-C-0591

United States Army Electronics Command, Department of Defense
Contract DAAB03-75-C-0419(L 433)

National Institute of Child Health and Human Development
Contract N01-HD-1-2420

National Institutes of Health
General Research Support Grant RR-5596

HASKINS LABORATORIES

Personnel in Speech Research

Alvin M. Liberman,* President and Research Director
Franklin S. Cooper, Associate Research Director
Patrick W. Nye, Associate Research Director
Raymond C. Huey, Treasurer
Alice Dadourian, Secretary

Investigators

Arthur S. Abramson*
Thomas Baer*
Peter Bailey[1]
Fredericka Bell-Berti*
Gloria J. Borden*
James E. Cutting*
Ruth S. Day*
Michael F. Dorman*
Frances J. Freeman*
Jane H. Gaitenby
Thomas J. Gay*
Terry Halwes
Katherine S. Harris*
Alice Healy*
Isabelle Y. Liberman*
Leigh Liske[1]
Ignatius G. Mattingly*
Paul Mermelstein
Seiji Niimi[2]
Lawrence J. Raphael*
Bruno H. Repp*
Philip E. Rubin*
Donald P. Shankweiler*
George N. Sholes
Michael Studdert-Kennedy*
Quentin Summerfield[1]
Michael T. Turvey*

Technical and Support Staff

Eric L. Andreasson
Dorie Baker*
Elizabeth P. Clark
Cecilia C. Dewey
Donald S. Hailey
Harriet G. Kass*
Sabina D. Koroluk
Christina R. LaColla
Roderick M. McGuire
Agnes McKeon
Terry F. Montlick
Loretta J. Reiss
William P. Scully
Richard S. Sharkany
Edward R. Wiley
David Zeichner

Students*

Mark J. Blechner
Steve Braddon
David Dechovitz
Susan Lea Donald
Donna Erickson
F. William Fischer
Hollis Fitch
Carol A. Fowler
Morey J. Kitzman
Gary Kuhn
Andrea G. Levitt
Roland Mandler
Leonard Mark
Robert F. Port
Sandra Prindle
Abigail Reilly
Robert Remez
Helen Simon
Emily Tobey
Harold Tzeutschler
James M. Vigorito

---

*Part-time
[1]Visiting from The Queen's University of Belfast, Northern Ireland.
[2]Visiting from University of Tokyo, Japan.

CONTENTS

I. Manuscripts and Extended Reports

Exploring the Relations between Reading and Speech -- Donald Shankweiler and Isabelle Y. Liberman 1

On Interpreting the Error Pattern in Beginning Reading -- Carol A. Fowler, Isabelle Y. Liberman, and Donald Shankweiler. 17

Comments on the Session: Perception and Production of Speech II; Conference on Origins and Evolution of Language and Speech -- A. M. Liberman 29

Consonant Environment Specifies Vowel Identity -- Winifred Strange, Robert R. Verbrugge, Donald P. Shankweiler, and Thomas R. Edman. 37

What Information Enables a Listener to Map a Talker's Vowel Space? -- Robert R. Verbrugge, Winifred Strange, Donald P. Shankweiler, and Thomas R. Edman. 63

Identification of Dichotic Fusions -- Bruno H. Repp. 95

Discrimination of Dichotic Fusions -- Bruno H. Repp. 123

Coperception: Two Further Preliminary Studies -- Bruno H. Repp. 141

"Posner's Paradigm" and Categorical Perception: A Negative Study -- Bruno H. Repp. 153

Weak Syllables in a Primitive Reading-Machine Algorithm -- George Sholes. 163

Control of Fundamental Frequency, Intensity, and Register of Phonation -- Thomas Baer, Thomas Gay, and Seiji Niimi. 175

The Effect of Delayed Auditory Feedback on Phonation: An Electromyographic Study -- M. F. Dorman, F. J. Freeman, and G. J. Borden 187

Some Aspects of Coarticulation -- Fredericka Bell-Berti and Katherine S. Harris. 197

The Function of Strap Muscles in Speech -- Donna Erickson and James E. Atkinson. 205

Laryngeal Muscle Activity in Stuttering -- Frances J. Freeman and Tatsujiro Ushijima 211

II. Publications and Reports 239

III. Appendix: DDC and ERIC numbers (SR-21/22 - SR-44) 241

I. MANUSCRIPTS AND EXTENDED REPORTS

Exploring the Relations between Reading and Speech*

Donald Shankweiler[+] and Isabelle Y. Liberman[+]

ABSTRACT

Acknowledgment of the priority of the spoken language and the derivative nature of the writing system is an essential starting point for an investigation of reading acquisition in children. The relations between the language and the writing system are manifold and complex so that spoken sounds and alphabetic characters cannot be related in a one-to-one fashion. There is reason to believe that the phonetic level of representation plays an especially significant role in the acquisition of reading in the young child. We considered that a primary function of a phonetic representation is to yield an adequate span in working memory to permit linguistic interpretation of the temporally arrayed segments of the message. Results of our studies of short-term memory in good and poor readers suggested that the poor reader is deficient in forming a phonetic representation from speech as well as from script. In order to learn to read an alphabetically written language, the child must have, in addition to a phonetically organized short-term memory, the ability to make explicit the phonemic segmentation of his own speech. The findings indicate that in contrast to the tacit appreciation of phonemic differences in ordinary language use, explicit knowledge of the phonemic level is difficult to attain. Many children lack phonemic awareness when they start to learn to read and this may be a cause of reading failure.

---

*To be published in Neuropsychology of Learning Disorders: Theoretical Approaches, ed. by R. M. Knights and D. K. Bakker. (Baltimore: University Park Press).

+Also University of Connecticut, Storrs.

Acknowledgment: This work reflects the joint efforts of several individuals. The data were obtained by F. W. Fischer, C. Fowler, L. Mark, and M. Zifcak, who also assumed responsibility for their tabulation and statistical analysis. We are also indebted to A. M. Liberman, who suggested the hypothesis concerning the functions of the phonetic representation. Full details of the experiments on phonetic coding in recall will be presented in a paper in preparation.

[HASKINS LABORATORIES: Status Report on Speech Research SR-45/46 (1976)]

Given so little agreement on how best to teach children to read, it is perhaps not surprising to find divergent conceptions of the nature of reading itself. Among these, we find two contrasting positions concerning the relationships between reading and speech. On the one hand, some writers (e.g., Goodman, 1968; Smith, 1973) have tended to ignore the relationship, choosing instead to emphasize the relative autonomy of reading and writing. Their counsel is, in effect, to forget about speech when teaching reading. A major target of their criticism has been the so-called phonic approach to reading instruction, which stresses the letter-to-sound mappings while failing to appreciate that we cannot read simply by concatenating individual letter sounds. On the other hand, we and a few other investigators (Huey, 1908; Mattingly, 1972; Shankweiler and Liberman, 1972; Rozin and Gleitman, in press) have emphasized the importance of the derivative nature of reading and writing and the intimate connection between speech and the alphabet. In defending this aspect of the study of reading, however, we give due weight to the complexity of the relationship. We believe that many of the criticisms that have been raised would apply only to a very simplistic view of how spoken sounds and alphabetic characters are related.

Central to the understanding of how reading is acquired, in our view, is the question of how reading builds on the speech processes of the child. We know, of course, that spoken language is historically prior to reading and writing in the development of the race, ontogenetically prior in the life of the individual, and logically prior to the relation of written symbols to their speech referents. Further evidence of the derivative status of writing and reading and the practical importance of the priority of speech is readily at hand. Consider the contrasting situations of the congenitally blind and the congenitally deaf. The blind acquire spoken language normally; the profoundly deaf, even under the most favorable conditions, are so effectively isolated from language that they show severe deficiencies in every aspect of language development (Furth, 1966). Since the blind child learns to read by means of the substitute sense of touch, we may ask why the deaf child cannot effectively exploit his intact visual channel for reading. Presumably he cannot do so because deafness blocks the development of a foundation in primary language so necessary as a basis for learning to read. If reading were, as some have argued, an alternative and coequal language reception system, then it would be hard to explain why the deaf could not learn language by eye as readily as the hearing learn it by ear. Our interest, of course, is to understand the acquisition of reading in children with intact sensory capacities. We make reference to reading in the blind and the deaf only to emphasize how closely reading is tied to speech.

If reading and speech are so closely linked, we would expect them to share much of the same neural machinery. As Halwes (1968) has pointed out, it is unparsimonious to imagine a completely parallel language understanding system (for reading) that borrowed nothing from the primary speech system. Rather than developing a separate device for reading, it would be more parsimonious to expect that the would-be reader modifies the speech perception system to accept optical information. We assume that the speech system works by mapping the acoustic signal into progressively more abstract representations, and we assume that the reading device must tie in with that system at some level. How much visual processing must be done before script can be represented in the common language processing system (as though the input had been speech rather than script)? To put the question another way, what is the level of representation at which script is recoded?

Certain facts about the writing system must constrain how we conceive of the reading process. All writing systems make contact at some point with the spoken language. Some, like Chinese and Japanese logographs, tie in at the level of words, others at the level of the syllable. Some--the alphabets--link their primary symbols to distinctive aspects of the sound structure of the language. In the case of English, there is good reason to believe that script makes contact with the primary language system at more than one level. At times, similarity of spelling may denote not similarity of sound, but similarities of origin and root meaning, as in such word pairs as sign and signal. Such cases are not uncommon. Moreover, the assignment of grammatical class is sometimes preliminary to determining the correct phonetic form. To use an example of Rozin and Gleitman (in press), the written word contract is ambiguous until we know whether it functions as a noun or as a verb. The correct phonetic representation of such ambiguous words cannot be fully attained without reference to more molar representations. These observations obviously constrain our choices when we attempt to model the perceptual system in reading. Thus, we do not assume that the reader is tied to a rigid hierarchy of successive processing stages. Rather, we suppose that the transformation of script into speech occurs at a number of levels concurrently and in parallel.

To recapitulate, the fundamental task of the beginning reader is to construct a link between speech and the arbitrary signs of script. Although the alphabet is roughly a cipher on the phonemes of speech, this does not imply that learning to read is merely a matter of acquiring letter-to-sound correspondence. English spelling does not fully reflect the phonetic facts of the language, and at times seems deliberately to ignore them in order to convey other kinds of information helpful to the easy comprehension of what is read. We assume that the experienced reader learns to detect and to exploit such multileveled representation, though the complexity of the orthography is surely an added source of difficulty for the beginning reader.

FUNCTIONS OF THE PHONETIC REPRESENTATION

Although English spelling is not a faithful phonetic transcription, there is reason to suppose that the phonetic level of representation plays an especially significant role in the acquisition of reading in the young child. Even in English the alphabet is largely keyed to the sound structure. Hence, new words can be given at least an approximate pronunciation on first encounter if the reader understands how the alphabet works. Obviously, the reader must recode phonetically if he is to obtain the phonetic realization of a new word. But what does he do with words and phrases he has read many times? Does he in these cases construct a phonetic representation, or does he, as some believe, bypass the phonetic level and go directly from visual shape to meaning?

It seems likely that phonetic recoding might occur even with frequently read materials, and that its persistence in older, more experienced readers is not to be regarded merely as a habit that has ceased to be functional. The possibility we are proposing is that the reader needs a phonetic base on which to extract the message from its encipherment in script; that is, the normal primary language processes of storing, indexing, and retrieving from the dictionary in our heads are carried out by means of a phonetic code. Moreover, in addition to the possibility that the dictionary may be indexed phonetically, consider what cues we use to decode the syntax of the message. Here we are aided by the rise and fall of the speech melody and its pattern of rhythms and

stresses. These are not given directly in script, and it may require the mediation of an internal phonetic representation to enable the reader to construct those prosodic features so necessary to comprehension (Liberman, Shankweiler, Liberman, Fowler, and Fischer, in press).

Since the perceiver cannot process each message unit fully at the time of its arrival, we may be sure that short-term memory is one of the primary linguistic processes essential to comprehension of both written and spoken language. The perceiver, whether functioning as reader or hearer, must hold a sufficient number of shorter segments (words) in memory in order to apprehend the longer segments (sentences). Obviously, if he had a span of only two words, the perceiver's comprehension of connected discourse would be extremely limited. But does the reader form a different kind of memory representation than the hearer? Although we do not rule out the possibility that read words can be held temporarily in some visual form, we have indicated reasons above for supposing that the reader typically engages in recoding from script to some phonetic form. [See Liberman, Mattingly, and Turvey (1972) for a fuller exploration of the suggestion that the phonetic representation is uniquely suited to the short-term storage requirements of language.]

Apart from these speculations, there is much relevant experimental evidence for phonetic recoding. In many investigations it has been found that when lists of letters or alphabetically written words are presented orthographically to be read and remembered, the confusions in short-term memory are based on phonetic rather than visual similarity (Sperling, 1963; Conrad, 1964, 1972; Baddeley, 1966, 1968, 1970; Hintzman, 1967; Kintsch and Buschke, 1969). From these findings, it has been inferred that the stimulus items had been stored in phonetic form rather than in visual form. Conrad (1972) has emphasized that the tendency to recode visually presented items into phonetic form is so strong that subjects do this even in experimental situations in which to do so penalizes recall.

There is evidence from a similar kind of experiment (Erickson, Mattingly, and Turvey, 1973) that phonetic recoding occurs even when the linguistic stimuli are not presented in an alphabetic form that represents the phonetic structure, but in a form (the Japanese kanji characters) that represents the semantic message more directly. Moreover, under some circumstances, even nonlinguistic stimuli may be recoded into phonetic form and stored in that form in short-term memory. In this connection, Conrad (1972) found that in recall of pictures of common objects, the confusions of children aged six and over were clearly based on the phonetic forms of the names of the objects rather than on their visual or semantic characteristics.

To be sure, none of these experiments dealt with wholly natural reading situations, since most involved the reading of isolated words and syllables rather than connected text. They are nevertheless relevant to the assumption that even the skilled reader might recode phonetically in order to gain an advantage in short-term memory and to utilize the primary language processes he already has available to him. It remains to be determined whether good and poor readers among children in the early states of reading acquisition are distinguished by greater or lesser tendencies toward phonetic recoding.

PHONETIC RECODING IN GOOD AND POOR BEGINNING READERS

In view of the short-term memory requirements of the reading task and evidence for the involvement of phonetic coding in short-term memory, we might expect to find that those beginning readers who are progressing well and those who are doing poorly will be further distinguished by the degree to which they rely on phonetic recoding.

In exploring this possibility, we studied three groups of school children nearing completion of the second year of elementary school who differed in level of reading achievement as measured by the word recognition subtest of the Wide Range Achievement Test (Jastak, Bijou, and Jastak, 1965). The first group, the superior readers, comprised 17 children reading about two years above their grade placement. The other two groups (whom we originally designated marginal and poor readers) can be considered together as the "poor readers" since their performances in these experiments were not significantly different from each other. Together the poor readers included 29 children averaging from one-half to a full year of reading retardation and roughly equated with the superior readers in mean age and IQ.

The experimental procedure was similar to one devised by Conrad (1972) in which the subject's performance is compared on recall of phonetically confusable (rhyming) and nonconfusable (nonrhyming) letters. Our expectation was that phonetically similar items would maximize phonetic confusability and thus penalize recall in subjects who use the phonetic code in short-term memory. Strings of five uppercase letters were presented tachistoscopically in a simultaneous 3-sec exposure. Half were composed of rhyming consonants (drawn from the set B C D G P T V Z) and half were composed of nonrhyming consonants (drawn from the set H K L Q R S W Y).

The test was given twice: first with immediate recall, then with delayed recall. In the first condition, recall was tested immediately after presentation by having subjects print as many letters as could be recalled in each letter string, in the order given. To make the task maximally sensitive to the recall strategy, we then imposed a 15-sec delay between tachistoscopic presentation and the response of writing down the string of letters. The children were requested to sit quietly during the delay interval; no intervening task was imposed. We have reason to believe that the subjects used this period for rehearsal, since many were observed mouthing the syllables silently.

The responses were scored in two ways, with and without regard to serial position. In the first scoring procedure, only those items listed in the correct serial position were counted correct. The second scoring procedure credited any items that occurred in the stimulus set regardless of the order in which they were written down. The pattern of results was remarkably similar, given data derived from each method of scoring. Ability to recall in correct serial order is apparently not the major factor that distinguishes good and poor readers on this task.

As was to be expected, the phonetic characteristics of the items influenced the rate of correct recall. This may be seen in Figure 1, which shows the results summed over serial positions. The circles give the error rates for strings of rhyming items (labeled "confusable"); the triangles give errors on recall of the nonrhyming ("nonconfusable") strings. In all groups, there were

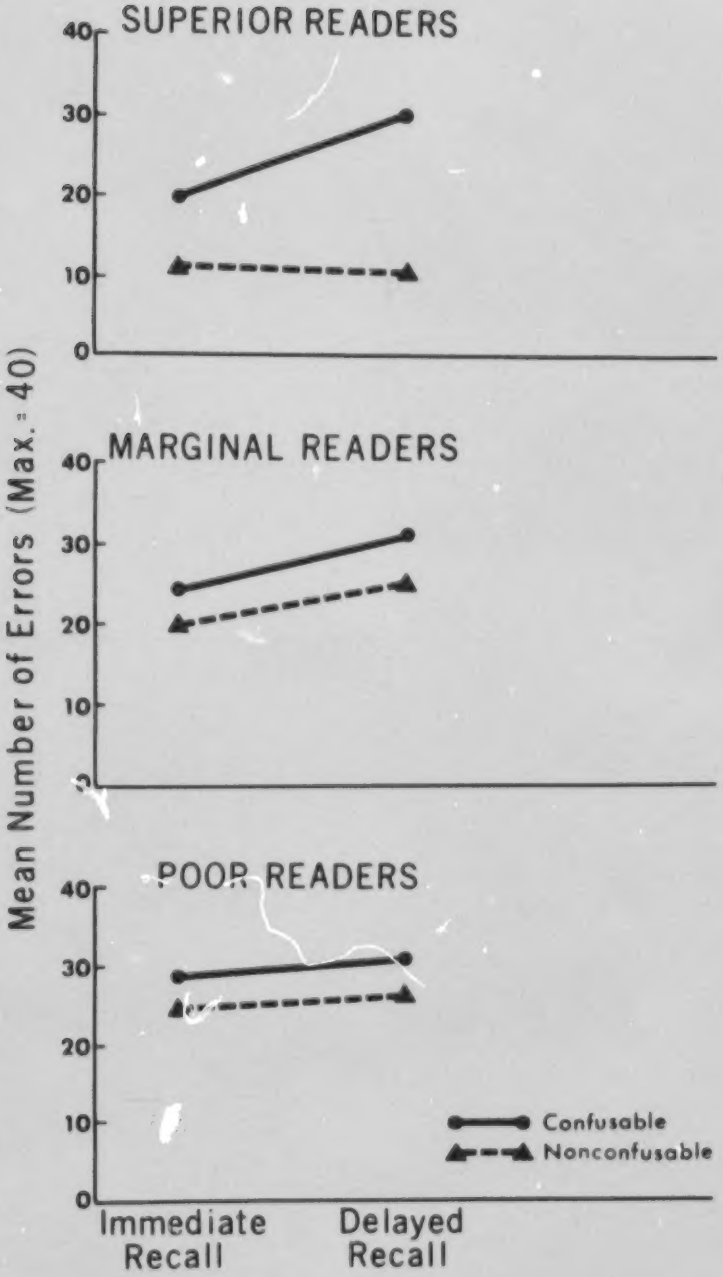


Figure 1: Mean recall errors summed over serial positions.

significantly more errors on recall of the confusable items. However, there were notable differences in the effects of phonetic similarity on the recall of children who differed in reading level. It is apparent from the figure that the main differences are between superior readers and the other groups.

The net effect of phonetic confusability on recall was much greater in the superior readers than in the others. It would be difficult to explain this result by assuming that the groups differ merely in general memory capacity. Superior readers were clearly better at recall on nonconfusable items than were the poor readers, while, at the same time, failing to show a clear advantage on the confusable items. We regard this as an interesting result. It is a relatively easy matter to demonstrate that poor readers do less well than good readers on a variety of language-dependent tasks. But here, by manipulation of the phonetic characteristics of the items, we have virtually eliminated the advantage of the superior readers.

As we said, recall was measured on half the trials immediately after presentation of the display, and on the other half after a 15-sec delay. Turning back to Figure 1, we see that delay magnified the penal effect of phonetic confusability, but only in the superior readers. Figure 2 shows plots of the error rates at each serial position. Viewing the results of the delay condition (shown in the lower portion), we see that the superior readers are sharply distinguished from the others in recall of nonconfusable items and nearly indistinguishable in their recall of confusable items. Why should imposing a delay between stimulation and recall affect good and poor readers differently? Is it simply the case that good readers try harder and rehearse the items more vigorously? Although we cannot be sure, we do not think that vigor or rate of rehearsal is a factor that chiefly distinguishes good and poor readers on this task. Certainly we know that the poor readers were attempting to rehearse because they so often mouthed the items during the interval (Liberman et al., in press).

We considered and rejected other explanations of the pattern of results obtained by good and poor readers. (1) The difference between the groups cannot easily be attributed to briefer memory span in the poor readers. Even if it were generally true that poor readers have briefer spans, the differential effect of phonetic similarity on recall performance by the two groups would still require explanation. (2) To suppose that the poor readers suffer mainly from a difficulty in reproducing the <u>order</u> of the items in the memory set encounters the same difficulty. Moreover, as we said, the pattern of results is much the same when the scoring credited the correct items in each string regardless of serial position.

The interpretation we find most plausible and interesting is that the results reflect genuine differences between good and poor readers in their use of a phonetic code. Of course we cannot argue that phonetic coding is entirely absent in the poor readers, since they demonstrated significant effects of confusability, though of lesser magnitude. A weak or defective phonetic representation in the poor readers could account for the failure of rehearsal to be effective.

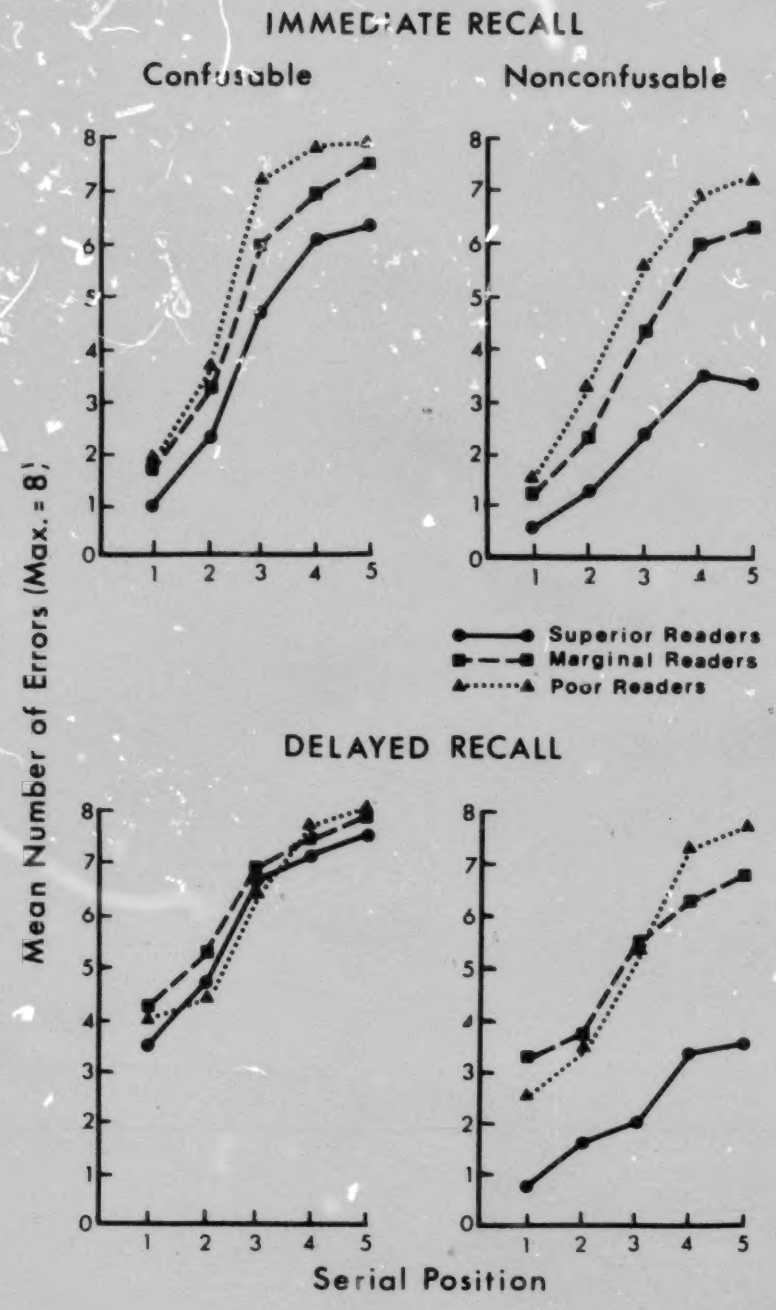


Figure 2: Recall data replotted as a function of serial position.

8

AN AUDITORY ANALOG AND ITS VISUAL COUNTERPART

In light of the foregoing results, it seemed reasonable to suppose that poor readers may have a specific difficulty in constructing a phonetic representation from script. Before we could accept this hypothesis, however, we needed to find out what would happen when confusable and nonconfusable items were presented by ear. Since phonetic coding is presumably inescapable when speech material arrives auditorily, presentation by ear should force the poor reader into a phonetic mode of information processing. If an important component of his difficulty is a deficiency in recoding visual symbolic material into phonetic form, then the phonetic similarity of auditorily presented rhyming items should affect him as much (or as little) as it does the superior readers. Quantitative differences in memory capacity between the two groups may still show up in the general level of recall on the auditory presentation, but the statistical interaction of reading level and phonetic confusability should be diminished. If, on the other hand, the interaction remained, then it would follow that the difference between good and poor readers in regard to the use of a phonetic representation is not specifically linked to the visual information channel.

Two new experiments were carried out on the same subjects in order to clarify this important point. Since auditory presentation requires successive input, a parallel experiment was designed with visual serial presentation. Except in minor details, the results are like those previously obtained for simultaneous presentation of the letters, and, to our surprise, the visual and auditory experiments differed hardly at all in their results. The findings of each experiment are displayed in Figure 3, which gives serial position curves for recall of auditorily presented and visually presented items. As in the earlier experiment, the performances of the groups representing the extremes of reading ability differed mainly on the phonetically dissimilar items. Once again, phonetic similarity produced a greater impact on the superior readers than on the poor ones. It made practically no difference to the results whether the items to be recalled were presented to the eye or to the ear. Apparently, the crux of the difficulty for the poor reader on these tasks cannot be pinpointed as specifically as we originally believed. Though poor readers may indeed experience difficulty in the transformation of visual features into phonetic ones, the root problem is more general.

These new experiments lead us to expect that differences between good and poor readers will turn on their ability to determine and use a phonetic representation and not merely on their ability to recode from script. We suspect that individual differences in the availability of phonetic recoding strategies on recall tasks may indicate limits of the reader's active awareness of those aspects of language structure to which the alphabet is most directly keyed. This is a possibility that we shall wish to explore directly. We turn now to those aspects of cognitive development that are most relevant to use of an alphabet.

WHAT A CHILD NEEDS TO "KNOW" IN ORDER TO USE AN ALPHABET TO FULL ADVANTAGE

The preliterate child brings to the task of learning to read considerable competence in his spoken language. Our concern is to discover what additional abilities he needs in order to become a reader. Bolinger (1968) places the problem of the learner and the teacher of reading in proper perspective:

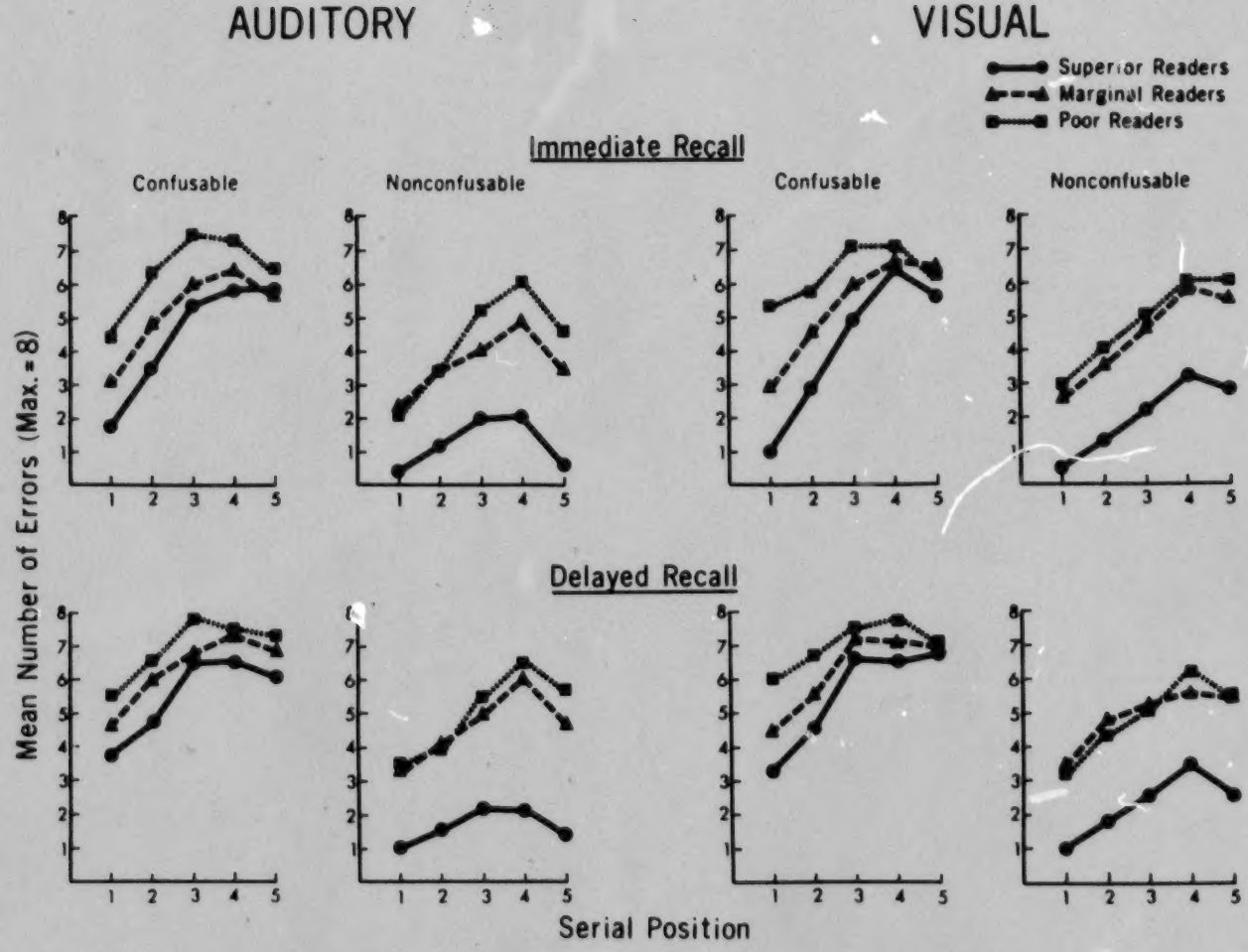


Figure 3: Recall data for the auditory analog experiment and its visual counterpart.

> When a child who is already almost fully equipped with a language comes to the task of reading, anything that will help him transfer what he already knows to what he is expected to write and read is priceless (p. 177).

We have argued that an efficient short-term memory system is a requirement for good comprehension of language, both by eye and by ear, and that this requirement is most efficiently met by a phonetic representation. Reading, however, poses an additional requirement. The child must also have ready conscious access to certain aspects of the contents of that memory; he must have, in Mattingly's (1972) phrase, a degree of "linguistic awareness." In order to realize fully the advantages of an alphabet, the user--child or adult--must know quite explicitly what speech segments are represented by the strings of letters (Liberman, 1973; Liberman et al., in press).

It is appropriate at this point to remind ourselves of the benefits that alphabets confer. As we have said, a unique advantage is that each new word does not have to be learned as if it were an ideographic character before it can be read. That is, given a word that is already in his mental word store, the reader can apprehend the word without specific instruction, though he has never seen it before in print; or, given a word that he has never before seen or heard, he can closely approximate its spoken form until its meaning can be inferred from context or discovered later by asking someone about it. By functioning, however roughly, as a surrogate for phonemes, the alphabet gives its users immediate access to all items in a vast word store by means of a highly economical symbol set.

The savings may be had, however, only by the user who knows how the alphabet works. As in all complex cognitive skills, alternative strategies are possible. The very diversity of the orthographies that have developed during the course of evolution of writing is testimony to the flexibility of the perceptual apparatus. It is possible to read words written by an alphabet as though they were logograms. Many children undoubtedly begin to read in this way. However, the unique advantages of the alphabet are closed to the child who cannot use it analytically; though he may translate the logograms into phonetic representation, this will not help him to apprehend new words. In order to make the alphabet work for him, the child has first to be able to make an explicit analysis of the segments of spoken language. He has to be able to analyze speech into words, syllables, and phonemes. The last mentioned is of particular importance for users of an alphabet, because the phoneme is the principal point at which the writing system meshes with the speech system.

When we speak of explicit knowledge of the segments in the spoken message, we wish to make it very clear that something more is involved than the ordinary competence required in language use. That is to say, a person may be a completely adequate speaker-hearer of his language without having the dimmest awareness that the spoken word bed contains three phoneme segments and bend contains four. The immediate recognition of these as different words, failing the ability to indicate that /n/ is the unshared segment, is an example of what Polanyi(1964) has called "tacit knowledge." Such knowledge is sufficient, or course, for comprehension of the spoken message. Writing and reading, on the other hand, demand an additional analytic capability. Even before the advent of writing, those who used speech poetically in songs and chant must have been able to count syllables in order to form the meter, and been aware of the phonemic level in

order to make rhymes. Some such explicit knowledge of these properties of speech is a precondition for understanding the alphabetic principle.

THE DIFFICULTIES OF MAKING SPEECH SEGMENTS EXPLICIT

Elsewhere (Liberman, 1971, 1973; Liberman, Shankweiler, Fischer, and Carter, 1974) we have considered why awareness of the phoneme might be rather difficult to attain. In brief, we referred to a fact about the acoustic structure of speech. Consonants and vowels are not discretely present in the signal, but are represented overlappingly in the syllable, a condition that has been called "encodedness" (Liberman, Cooper, Shankweiler, and Studdert-Kennedy, 1967). As a consequence, the word dig, for example, has three phonetic segments but only one acoustic segment. Analyzing an utterance into syllables, on the other hand, may present a different and easier problem. We expect this to be so because in most cases each syllable has a distinctive peak in acoustic energy. The cue of auditory amplitude is a crude one that could not be used to locate exact syllable boundaries, but it can serve to indicate to the listener how many syllables there are in an utterance.

The merging of phones in the sound stream complicates the process of discovery of the phonemic level of speech for the would-be reader. This is not to say, of course, that the young child has difficulty differentiating word pairs, such as bad and bat, that differ in only one phoneme. There is evidence (Read, 1971) that children hear these differences quite as accurately as adults. The problem is not, as many believe, to get the child to discriminate such word pairs, but rather to lead him to appreciate that each of these words contains three segments, and that they are alike in the first two and differ in the third. This is a further example of the distinction we drew earlier between tacit and explicit knowledge of the phonetic structure of language.

The encoded nature of the phonemes has another consequence that surely contributes to the difficulty of learning to read analytically. It makes it impossible to read by sounding out the letters one by one. In the example of dig, used above, reading letter by letter gives, not "dig," but "duhiguh." In order to learn to read analytically, one must instead discover how many of the letter segments must be taken simultaneously into account in order to arrive at the correct phonetic rendition. In the case of the word dig, there is reason to believe the number would be three. But, in fact, there is no simple rule for arriving at that number, and we suspect that learning to group the letters for the purpose of proper phonetic recoding is one of the really significant skills one must acquire. Thus, even in languages such as Finnish and Spanish in which the writing system closely approximates one-to-one correspondences between letters and phonemes, reading cannot be a simple matter of association between alphabetic characters and spoken sounds. In order to recover the spoken form, the reader must still "chunk" all the letters that represent the phonetic segments encoded into each syllable. In the case of reading a word in isolation, the coding unit is probably the syllable. In reading connected text, the number of letters that must be apprehended before recovery of the spoken form may at times be quite large, for reasons we have discussed. We do not know how the coding unit may vary with the prosody of the text and the reader's experience, but we may be sure that such units almost always exceed one letter in length. Therefore, we would stress that making analytic use of an alphabet does not mean reading letter-by-letter.

The foregoing discussion has stressed that explicit awareness of the phonetic structure of utterances is a very different thing from the ability to distinguish words whose phonetic structure differs minimally. The latter is easy for every normal child of school age, whereas the difficulty of explicit analysis has been noted by a number of researchers (Bloomfield, 1942; Rosner and Simon, 1971; Calfee, Chapman, and Venezky, 1972; Savin, 1972; Elkonin, 1973; Gleitman and Rozin, 1973). However, there had been no experiments designed to demonstrate directly that phonetic segmentation is more difficult for young children than syllabic segmentation, and that the ability to do it might develop later.

DEVELOPMENT OF THE AWARENESS OF SPEECH SEGMENTS IN THE YOUNG CHILD

Recently, we (Liberman et al., 1974) investigated the development of the ability to analyze words explicitly in syllables and phonemes. The task was posed to the child subjects in the guise of a tapping game, in which segments had to be indicated by the number of taps. We found steep age trends for analysis of words into each kind of segment, but, at each age, test words were more readily segmented into syllables than into phonemes. At age four, none of the children in our sample could segment by phoneme (according to the criterion we adopted), while nearly 50 percent could segment by syllable. Even at age six, only 70 percent succeeded in phoneme segmentation, whereas 90 percent were successful in the syllable task.

Further research is needed to confirm and generalize these results. Since the syllable is also the unit of metric scan, it is conceivable that the motor response of tapping is more compatible with analysis by syllable than with analysis by phoneme. An alternative procedure, designed by Goldstein (1974), asks the child to indicate the number of segments in test words by counting out tokens, thus limiting rhythmic motor responses that might bias the outcome in favor of the syllable. Goldstein's preliminary work with this alternative procedure confirmed that phoneme segmentation is genuinely more difficult than syllable segmentation.

We hope eventually to clarify the meaning of the age trends we found. On the one hand, the increase in ability to segment phonetically might result from the reading instruction that typically begins between ages five and six. Alternatively, it might be a manifestation of cognitive growth not specifically dependent on training. The latter possibility could be tested by a developmental study of segmentation skills in a language community such as the Chinese, where the orthographic unit is the word and where reading instruction therefore does not demand the kind of phonetic analysis needed in an alphabetic system.

SEGMENTATION AND READING ACQUISITION

There is some evidence that the difficulties of phoneme segmentation may be related to problems of early reading acquisition. Such a relation can be inferred from the observation that children who are resistant to early reading instruction have problems even with spoken language when they are required to perform tasks demanding some rather explicit understanding of phonetic structure. Such children are reported (Monroe, 1932; Savin, 1972) to be deficient in rhyming, in recognizing that two different monosyllables may share the same first

(or last) phoneme segment, and also in playing certain speech games, which require a shift of the initial consonant segment of a word to a nonsense syllable suffix.

In our segmentation experiment, we noted a sharp increase in the number of children passing the phoneme-segmentation task, from only 17 percent at age five to 70 percent at age six. Hence, the steepest rise in segmentation ability coincides with the first intensive concentration on reading-related skills in the schooling of the child. This result, together with the observations on the lack of "transparency" of the phoneme to which we referred earlier, suggests a connection between phonetic segmentation ability and early reading acquisition. In a pilot study, we have begun to explore this relation. We measured the reading achievement of the children who had taken part in our experiment on phonemic segmentation described above. Testing at the beginning of the second school year, we found that half the children in the lowest third of the class in reading achievement--as measured by the word-recognition task of the Wide Range Achievement Test (Jastak et al., 1965)--had failed the phoneme segmentation task the previous June; on the other hand, there were no failures in phoneme segmentation among the children who scored in the top third in reading ability.

We are hopeful that studies of preschool children's ability to segment speech may shed some light on the matter of reading readiness. We plan to examine the pattern of reading errors in children at different levels of reading ability in relation to their ability to indicate the segments of spoken speech. If the indications of our pilot work are borne out, failure on both the syllable and the phoneme tasks at the first-grade level will be prognostic of extreme reading difficulty.

SUMMARY AND CONCLUSIONS

We believe the priority of spoken language and the derivative nature of reading and writing are the starting points for any understanding of the nature of writings systems and their acquisition. Reading, however, presents special problems for the perceiver, the nature of which reflects the manner in which the writing system makes contact with the primary speech system. In the case of English, the ties between the language and its spelling are based only partly on the sound structure. Nevertheless, it is particularly appropriate to direct the child's attention to the phonemic level, because the phonemic correspondences are the entry points to any alphabetic writing system.

We considered that a primary function of a phonetic representation, whether for the listener or the reader, is to yield an adequate span in working memory to permit linguistic interpretation of the temporally arrayed segments of the message. Results of our studies of short-term memory in good and poor readers suggested that the poor reader is deficient in forming a phonetic representation from speech as well as from script.

In order to learn to read an alphabetically written language, the availability of a phonetically organized short-term memory is not sufficient. In addition, the child must have the ability to make explicit the segmentation of his own speech, particularly at the level of the phoneme. Data were presented indicating that explicit knowledge of the phonetic level is difficult to attain in contrast to the tacit appreciation of phonemic differences reflected in

ordinary language use. We and others have noted that phonemic awareness is lacking in many children when they start to learn to read, and may be a cause of reading failure. In sum, the relations between speech and reading are both intimate and subtle. It would seem appropriate for the early instruction in reading to place initial stress on making the child aware of the speech segments he will eventually learn to represent by written signs.

REFERENCES

Baddeley, A. D. (1966) Short-term memory for word sequences as a function of acoustic, semantic, and formal similarity. Quart. J. Exp. Psychol. 18, 362-365.

Baddeley, A. D. (1968) How does acoustic similarity influence short-term memory? Quart. J. Exp. Psychol. 20, 249-264.

Baddeley, A. D. (1970) Effects of acoustic and semantic similarity on short-term paired associate learning. Brit. J. Psychol. 61, 335-343.

Bloomfield, L. (1942) Linguistics and reading. In Elementary English 18, 125-130; 18, 183-186.

Bolinger, D. (1968) Aspects of Language. (New York: Harcourt, Brace & World).

Calfee, R., R. Chapman, and R. Venezky. (1972) How a child needs to think to learn to read. In Cognition in Learning and Memory, ed. by L. W. Gregg. (New York: Wiley).

Conrad, R. (1964) Acoustic confusions in immediate memory. Brit. J. Psychol. 55, 75-84.

Conrad, R. (1972) Speech and reading. In Language by Ear and by Eye: The Relationships between Speech and Reading, ed. by J. F. Kavanagh and I. G. Mattingly. (Cambridge, Mass.: MIT Press).

Elkonin, D. B. (1973) U.S.S.R. In Comparative Reading, ed. by J. Downing. (New York: Macmillan).

Erickson, D., I. G. Mattingly, and M. T. Turvey. (1973) Phonetic activity in reading: An experiment with kanji. Haskins Laboratories Status Report on Speech Research SR-33, 137-156.

Furth, H. (1966) Thinking without Language: Psychological Implications of Deafness. (New York: The Free Press).

Gleitman, L. R. and P. Rozin. (1973) Teaching reading by use of a syllabary. Read. Res. Quart. 8, 447-483.

Goldstein, D. M. (1974) Learning to read and developmental changes in covert speech and in a word analysis and synthesis skill. Unpublished Ph.D. dissertation, University of Connecticut.

Goodman, K. S. (1968) The psycholinguistic nature of the reading process. In The Psycholinguistic Nature of the Reading Process, ed. by K. S. Goodman. (Detroit: Wayne State University Press).

Halwes, T. (1968) Comment. In Communicating by Language, ed. by J. F. Kavanagh. (Bethesda, Md.: NICHD), p. 160.

Hintzman, D. L. (1967) Articulatory coding in short-term memory. J. Verbal Learn. Verbal Behav. 6, 312-316.

Huey, E. B. (1908) The Psychology and Pedogogy of Reading. (New York: Macmillan).

Jastak, J., S. W. Bijou, and S. R. Jastak. (1965) Wide Range Achievement Test. (Wilmington, Del.: Guidance Associates).

Kintsch, W. and H. Buschke. (1969) Homophones and synonyms in short-term memory. J. Exp. Psychol. 80, 403-407.

Liberman, A. M., F. S. Cooper, D. Shankweiler, and M. Studdert-Kennedy. (1967) Perception of the speech code. Psychol. Rev. 74, 431-461.

Liberman, A. M., I. G. Mattingly, and M. T. Turvey. (1972) Language codes and memory codes. In Coding Processes in Human Memory, ed. by A. W. Melton and E. Martin. (Washington, D.C.: V. H. Winston & Sons).

Liberman, I. Y. (1971) Basic research in speech and lateralization of language: Some implications for reading disability. Bull. Orton Soc. 21, 71-87.

Liberman, I. Y. (1973) Segmentation of the spoken word and reading acquisition. Bull. Orton Soc. 23, 65-77.

Liberman, I. Y., D. Shankweiler, F. W. Fischer, and B. Carter. (1974) Reading and the awareness of linguistic segments. J. Exp. Child Psychol. 18, 201-212.

Liberman, I. Y., D. Shankweiler, A. M. Liberman, C. Fowler, and F. W. Fischer. (in press) Phonetic segmentation and recoding in the beginning reader. In Reading: Theory and Practice, ed. by A. S. Reber and D. Scarborough. (Hillsdale, N. J.: Lawrence Erlbaum Assoc.).

Mattingly, I. G. (1972) Reading: The linguistic process and linguistic awareness. In Language by Ear and by Eye: The Relationships between Speech and Reading, ed. by J. F. Kavanagh and I. G. Mattingly. (Cambridge, Mass.: MIT Press).

Monroe, M. (1932) Children Who Cannot Read. (Chicago: University of Chicago Press).

Polanyi, M. (1964) Personal Knowledge: Towards a Post-Critical Philosophy. (New York: Harper & Row).

Read, C. (1971) Pre-school children's knowledge of English phonology. Harvard Educ. Rev. 41, 1-34.

Rosner, J. and D. P. Simon. (1971) The auditory analysis test: An initial report. J. Learn. Dis. 4, 40-48.

Rozin, P. and L. R. Gleitman. (in press) The structure and acquisition of reading. In Reading: Theory and Practice, ed. by A. S. Reber and D. Scarborough. (Hillsdale, N. J.: Lawrence Erlbaum Assoc.).

Savin, H. B. (1972) What the child knows about speech when he starts to learn to read. In Language by Ear and by Eye: The Relationships between Speech and Reading, ed. by J. F. Kavanagh and I. G. Mattingly. (Cambridge, Mass.: MIT Press).

Shankweiler, D. and I. Y. Liberman. (1972) Misreading: A search for causes. In Language by Ear and by Eye: The Relationships between Speech and Reading, ed. by J. F. Kavanagh and I. G. Mattingly. (Cambridge, Mass.: MIT Press).

Smith, F. (1973) Psycholinguistics and Reading. (New York: Holt, Rinehart & Winston).

Sperling, G. (1963) A model for visual memory tasks. Human Factors 5, 19-31.

On Interpreting the Error Pattern in Beginning Reading

Carol A. Fowler,* Isabelle Y. Liberman,* and Donald Shankweiler*

ABSTRACT

The error pattern in beginning reading was examined from two perspectives: the location of a misread consonant or vowel segment within the syllable and the phonetic relationship between a consonant or vowel and a misreading of it. The first analysis showed, as earlier work had led us to expect, that consonants in the final position in a syllable were more frequently misread than initial consonants. In contrast, the position of a vowel within the syllable had no effect on the frequency with which it was misread. With regard to the second analysis, consonant errors were found to bear a close phonetic relationship to their target sounds, while errors on vowels were essentially unrelated, phonetically, to the vowel as written. The striking differences, demonstrated by the results of both analyses, between the consonants and the vowels were attributed to the different linguistic functions of the two types of segments and to their different representations in English orthography. These findings underscore the importance of nonvisual, language-related cognitive operations in reading acquisition.

By analyzing the errors that children make when they read, we can expect to learn something about the underlying difficulties of reading acquisition. However, analysis of beginners' errors can be enlightening only if the errors form patterns, and then only if we can make sense of the patterns in terms of what we know about those processes of language and perception on which the development of reading must depend. Of course, patterns do not reveal themselves automatically. Suitable strategies for examining the errors must be chosen by the investigator, and these naturally reflect one's views of the nature of the problem. That is to say, the choice of strategies for analysis of misreadings reflects our expectations and biases concerning what it is that makes learning to read difficult.

It seems patent to us that many children who lag behind in reading acquisition do not understand the nature of the link between the writing system and the language they already command in speech. Our research has therefore been directed to the problems the child encounters in mapping the letter signs of the written word to the linguistic segments of the spoken word. For this purpose, we have chosen to focus on the child's error pattern in reading isolated words rather than his reading of words in connected text. Our major reason for

---

*Also University of Connecticut, Storrs.

[HASKINS LABORATORIES: Status Report on Speech Research SR-45/46 (1976)]

adopting this approach was a practical one: it is more feasible to assess a child's analytic knowledge of the writing system when the materials used are as free as they can be from the contextual cues supplied by ordinary meaningful discourse. Empirical support for the validity of this approach is provided by earlier studies (Shankweiler and Liberman, 1972) in which we found a high correlation between children's ability to decode isolated words and their ability to read meaningful, connected text with comprehension.

Given the word as the unit for investigation, our strategy was to examine the beginner's misreadings from two perspectives: the location within the word where errors most frequently occur, and the phonetic relationships between the word as written and the child's incorrect renditions.

PHONOLOGICAL SEGMENTATION AND ERRORS IN BEGINNING READING

The first perspective was suggested to us by the results of an earlier experiment (Shankweiler and Liberman, 1972). In that experiment, we observed that the errors made by beginning readers did, in fact, show a pattern with respect to location within the word. Thus we noted, as others had (Daniels and Diack, 1956; Weber, 1970), that errors on final consonants far exceed those on initial consonants in a consonant-vowel-consonant (CVC) syllable. Additionally, we found that errors on medial vowels exceed errors on consonants in both the initial and final positions.

To account for this observed distribution of errors, we adopted a line of reasoning previously suggested by one of us (Liberman, 1971, 1973) in which it was argued that if the child is to take full advantage of an alphabetic writing system, he must be able to segment the spoken word into its component phonological units. That is to say, he first has to recognize that the continuous acoustic signal that constitutes the spoken word may be represented as an ordered string of discrete phonological segments. Second, the child must be able to identify explicitly the set of phonological segments that makes up a given word. Only by so doing can he acquire and use the orthographic rules that map these abstract units of sound onto their appropriate graphic representations. It is not enough that the child merely be able to discriminate words, such as bag and bat, which differ in one phoneme. Every normal child can do that long before attaining reading age. In order to learn to use an alphabet effectively, more is required than the perception of phonological differences. The child needs to know explicitly that, in the example given, the words each contain three segments and that they are alike in the first and second segments and differ in the third (cf. Gibson and Levin, 1975 and Rozin and Gleitman, in press, for extended discussions of the view).

Several recent investigations (Rosner and Simon, 1970; Calfee, Chapman, and Venezky, 1972; Liberman, 1973; Liberman, Shankweiler, Fischer, and Carter, 1974) of the phonological skills of young children have shown that many do indeed find the task of segmenting the spoken word a difficult one. In our study (Liberman et al., 1974), children in three age groups (nursery-preschool, kindergarten, and first grade) were asked to indicate, by means of a tapping game we showed them, the number of phonemes contained in each of a group of high-frequency words. Most of the youngest children were unable to perform the task as were the majority of the kindergarteners. Even at the end of the first grade, 30 percent of the children failed. The first-grade children who failed in the

segmentation task had considerably more difficulty later in reading acquisition than those who succeeded (Liberman, 1973).

In the light of these findings, it seemed reasonable to suppose that the task of phonological segmentation might also vary in difficulty with the position of a given segment in the syllable. That is, the initial sound in a syllable should be easiest to isolate for the purpose of relating sound to orthographic representation because it can be extracted without extensive analysis of the syllable's sound structure. Obversely, the final segment would be more difficult because just such an analysis would be required. The medial sound might be the most difficult to analyze because it is entirely embedded within the syllable. A report by Rosner and Simon (1970) seems to support these conjectures: when a child is asked to reproduce an auditorily presented word, but to leave out a specified consonant sound, he experiences the greatest difficulty with the medial consonant sound and the least difficulty with the initial sound.

One way to account for the error pattern observed in our earlier experiment, then, is to consider that it reflects the differential difficulty that the beginning reader experiences in segmenting sounds occurring in the initial, medial, and final positions in the syllable. Such an account would attribute the error difference we obtained between medial vowels, final consonants, and initial consonants to the relative positions within the syllable occupied by the different types of sounds and not to differences among the sound-types themselves.

Although the data of our previous experiment (Shankweiler and Liberman, 1972) are consistent with such an interpretation, controls were lacking that would enable us to rule out other possible interpretations. An adequate test of the hypothesis would require first that the set of consonants occurring in syllable-initial position be identical with the set that occurs in syllable-final position. Additionally, it would require that the vowel also occur in initial and final position, not only in the medial position, as was the case in our earlier experiment. If, in a test designed to incorporate these controls, errors on initial, medial, and final segments again rank as before, then we can conclude with more assurance that the order of difficulty reflects a true position effect for both consonants and vowels.

Accordingly, for the present experiment, we developed two word lists designed to meet these requirements.[1] In List 1, the 19 consonant phonemes that can occur in both the initial and in the final positions of a word appeared twice in each position.[2] In List 2, the seven vowel phonemes that can occur in the initial, medial, and final segment positions in a monosyllable appeared three times in each position. The items composing both the vowel and the consonant lists were monosyllabic words, which insofar as possible were familiar to third-graders (Buckingham and Dolch, 1936) but were not "sight" words.

---

[1] Ideally, it would have been desirable to provide both the consonant and the vowel controls within one list. Contingencies relating to reading and vocabulary level made this impossible to achieve.

[2] Medial consonants were excluded from the test list, as they had been in the earlier experiment. Their inclusion would restrict us to a very small set of consonants unless we allowed disyllables. Disyllables were avoided because we did not wish to introduce problems of syllable segmentation into the reading task.

The lists were presented in a single session.[3] The order of list presentation was balanced across subjects, and the order of words in each list was randomized. The test words were printed with a black felt-tip pen on separate unlined 3 × 5 file cards. The cards were placed face down in front of the subject and were turned over one by one by the examiner. The subject was asked to read each word as it was presented and to give his best guess if he did not know the word. Responses were phonemically transcribed by the examiner and were recorded on magnetic tape for later checking.

The subjects were children of the second, third, and fourth grades, 20 from each grade, chosen alphabetically from the rosters of male and female students in a public elementary school in Andover, Connecticut. Testing was done in the late fall and early winter.

THE SEGMENT POSITION EFFECT IN CONSONANT ERRORS

The distribution of phoneme frequencies in English is not the same in syllable-initial and in syllable-final segment positions. In order to control for possible effects of this difference, List 1 was constructed so that the same set of consonant phonemes appeared in each position. Despite this control, the error difference obtained in our earlier experiment was replicated. As can be seen in Table 1, final-consonant (FC) errors continue to exceed initial-consonant (IC) errors. The direction of the difference is the same at every grade level, and is consistent with the predicted rank ordering of difficulty of the initial and final segments in the syllable.

TABLE 1: Errors on initial and final consonants and on medial vowels (List 1) presented as proportions of opportunity for error (decimal points omitted).

Grade	IC	FC	MV[a]
2	08	16	27
3	05	10	15
4	02	06	08

[a]Occurrences not controlled.

An analysis of variance performed on the data indicated that the effect of consonant position was highly significant [$F(1,57) = 44.80$, $p < .001$]. As expected, there was also an increase in performance level with grade [$F(2,57) = 4.10$, $p < .025$]. The grade-by-position interaction was not significant.

Although the identity of the phonemes occurring in each segment position of the words was controlled in List 1, their orthographic representations were not controlled. Therefore, a further analysis was performed to ascertain that the larger FC error rates could not be ascribed to differences in the frequency or ease of apprehension of the different sets of orthographic representations that

[3]A third list, used to study orthographic complexity, was presented at the same time. It will be described in a later paper.

occur in the initial and final positions. For the purposes of this analysis, orthographic complexity was defined in two ways. First, it was defined in terms of the number of possible orthographic representations per phoneme. In this sense, a phoneme that can be spelled in many ways is more complex than one with few orthographic representations. Second, complexity was defined in terms of the number of letters in each orthographic representation. For example, "tch" would be more complex than "c." For the purposes of the following analysis, both criteria were used--that is, a phoneme was considered orthographically complex if it could be spelled in more than one way, but it also was considered complex if its single orthographic representation consisted of more than one letter. Based on these criteria, the consonant phonemes were separated into "simple" and "complex" categories.

In Table 2, IC and FC errors in the simple and complex categories are presented as proportions of opportunities for error. If orthographic complexity were the basis for the FC/IC difference, removing these phonemes on which FCs and ICs differ with respect to orthographic complexity should equalize the error rates. However, the difference is present even in the "simple" category whose member phonemes are simple both in syllable-initial and syllable-final position with respect to the indicated criteria.

TABLE 2: Errors on orthographically complex and simple sounds.[a] Errors presented as proportions of opportunity for error (decimal points omitted).

	Complex		Simple	
Grade	IC	FC	IC	FC
2	09	24	06	08
3	06	13	03	07
4	03	09	01	03

[a]Complex: /f,j,k,m,s,θ,ð,č,š,z/; simple: /b,d,g,l,n,p,t,r,v/.

Apparently then, neither phonemic distribution within the syllable nor orthographic complexity can account for the FC/IC difference in error rate. The difference, therefore, must be truly a position effect, that is, an effect of the location of a given phoneme in the syllable. Final-consonant segments are more difficult than IC segments because they are in the syllable-final position.

THE SEGMENT POSITION EFFECT IN VOWEL ERRORS

It can be seen from Table 3, which displays the error scores for the vowel-controlled list of words, that the vowels do not show the marked position effect of the consonants. The analysis of variance revealed only a marginally significant effect of segment position [$F(2,114) = 4.61, p < .05$]. Again, there was an increase in performance level with grade [$F(2,114) = 11.08, p < .01$], and the interaction was nonsignificant.

Analyses performed separately on the error scores for each grade show that the position effect for vowels was statistically significant at one grade level: the fourth grade. This is in contrast to the position effect for consonants,

TABLE 3: Errors on initial, medial, and final vowels and on initial and final consonants (List 2) presented as proportions of opportunity for error (decimal points omitted).

Grade	IV	MV	FV	IC[a]	FC[a]
2	47	43	43	17	32
3	28	27	31	09	19
4	20	12	19	04	11

[a]Occurrences not controlled.

which was significant in all three grades. Post-hoc means tests of the fourth-grade vowel data indicate that two differences accounted for the significant F values: errors on vowels in the initial position and in the initial and final positions combined, both significantly exceeded errors on vowels in the medial position. Thus, if a segment position effect for vowels can be said to exist at all, it must be attributed to the significantly fewer errors on medial than on initial and final vowels (and then only for the fourth-grade subjects).

THE RANK ORDERING OF CONSONANT ERRORS AND VOWEL ERRORS

We can now reexamine the vowel>final-consonant>initial-consonant rank ordering of errors that we observed in our original experiment. It should first be noted that because both the consonants and vowels could not be controlled within a single list, the consonant-vowel error hierarchy cannot be directly examined within either List 1 or 2. However, as can be seen in Tables 1 and 3, if vowel errors are scored in the consonant-controlled list and consonant errors in the vowel-controlled list, the vowel>final-consonant>initial-consonant hierarchy of error frequency is replicated at every grade level within both lists. It is clear that whereas vowels in any position elicit more errors than consonants, the initial-final difference among the consonants is maintained.

On the consonant-controlled list of the present experiment, the difference in error rate between the final consonants and the initial consonants found earlier was replicated even after phonological and orthographic differences between the two categories had been removed. The discrepancy, then, may be attributed to some difference in difficulty between the initial and final segment positions of the consonants in the syllable and not to the particular consonant phonemes or the orthographic patterns that tend to occur in the two syllabic locations.

On the other hand, the preponderance of vowel over consonant errors obtained in our earlier experiment can no longer be attributed to the embedded position of the vowel within the syllable. The results obtained with List 2 indicate that vowels are approximately equal in difficulty across the three syllable locations. We may conclude, therefore, that the vowels in our earlier experiment were more difficult than the consonants for the beginning readers, not because of their embedded location within the syllable, but, rather, because of characteristics specific to vowels and not present in consonants.

In summary, we have looked to see where the errors are made in the syllable and have concluded that there is a position effect for the consonants.

Syllable-final consonants give rise to twice as many errors as syllable-initial consonants. The position-related errors can therefore be viewed as an outcome of the difficulties of phonological segmentation. However, the frequency of vowel errors was not affected by the position of the vowel segment within the syllable. Therefore, we cannot regard the child's difficulties with vowels as a reflection of his inability to segment the syllable.

It may be argued (Liberman, 1973) that if the child's segmentation skills were improved, his difficulties with the vowels would not be a severe handicap to him in deciphering the text. This might be expected because the consonants carry the major information load in the word. If the child were able to assign correct sounds to the consonants in proper sequence, an incorrect rendition of the vowels would be corrected fairly easily in context.

THE NATURE OF THE PHONETIC ERRORS IN BEGINNING READING

Having considered the location of the errors, we turn our attention now to an examination of their nature. We found both in this experiment and a previous one (Shankweiler and Liberman, 1972) that vowels generate more errors than consonants. It is appropriate to ask how the errors might be different in the two phonetic classes. Our purpose in the following analysis was to look for phonetic relationships between the misread segment and the target segment. Of course, in ordinary reading the lexical and broader linguistic context may affect the choice of the guessed-at word. We deliberately minimized the contribution of context, as we have said, in order to be able to assign a relatively unambiguous interpretation to the errors that occur.

Because the experiment required the children to read the words aloud, all of them presumably had to make a transformation from a visual to a phonetic representation. We may be sure then that the child is recoding the material phonetically as he reads, and we can examine, segment by segment, the phonetic relationship between the child's misreading and the segment that would be produced in that position if the word were read as written. In order to make the examination, we have adapted techniques used by other investigators to examine errors of speech perception. There is much evidence from investigations of speech perception (see, for example, Miller and Nicely, 1955) that phoneme segments are themselves compounds of a small set of phonetic features and that errors in perceiving speech by ear can be understood on a feature basis. That is, a substituted phoneme, more often than not, is only a partial error, in the sense that it preserves features in common with the presented segment.

Recent data obtained by Eimas (in press) show that the pattern of consonant errors made by six- and seven-year-old children in recall of strings of visually presented nonsense syllables resembles extremely closely the pattern obtained with auditory presentation. Errors having more than one distinctive feature in common with the presented phoneme occurred significantly more frequently than errors sharing one or no features with the presented phoneme. These findings would lead us to expect that as the child reads, he recodes the input into a form that can be described in terms of a phonetic feature matrix.

If errors arise in the transformation from print to a phonetic code, then the pattern of errors due to misreading might be expected to resemble that due to mishearing. Thus, there is reason to expect that the frequency of misreading

would vary directly with the number of features shared between the presented and the misread segments. Factors other than degree of phonetic contrast, however, are likely to be involved in the misreadings of vowels. Whereas the rules relating spelling to phonetic segment are relatively straightforward for consonants, they are quite complex for vowels. For this reason, we might expect to find not only that more errors occur on vowels than on consonants, but also that the nature of the substitutions may be different for the two phonetic classes.

FEATURE SUBSTITUTION ERRORS AMONG CONSONANTS

To determine whether the misreadings among consonants pattern nonrandomly, we needed a way to quantify the phonetic distance between any two consonants. We also needed a way of comparing the observed frequency of errors at a given phonetic distance from a target phoneme with the frequencies that would be expected if the children were randomly assigning phonemes to letters.

For the purposes of this investigation, we defined phonetic distance in terms of the number of distinctive features shared by an error response and a target phoneme. Three features--voicing, place of articulation, and manner of articulation--describe the English consonants adequately, providing each with a unique feature description. For example, since /b/ and /p/ share two features, they are considered phonetically similar; /b/ and /s/, which share no features, are dissimilar. Each error response was classified in this manner, according to the number of features it shared with its respective target phoneme. The frequency of error responses in each of the phonetic-distance categories (zero, one, or two features shared) was tallied separately for children of each grade and for each consonant position.

Frequencies expected by chance were calculated by constructing a 19 × 22 triangular matrix with the 19 target phonemes (that is, the 19 consonants that appeared in List 1) represented vertically and the complete set of the 22 consonants of English represented horizontally. Each cell of the triangular matrix thus uniquely represented a target phoneme paired with a possible error response. We made the assumption that a child responding randomly would choose his responses only from among the set of English consonants. In each cell were listed the features shared in common by the appropriate target phoneme and error response. The frequencies of cells with entries containing zero, one, or two features shared by the target consonant and each possible erroneous response were tallied separately. These were expressed as proportions of errors that would be expected to share zero, one, or two features with the target phoneme if the children were assigning phoneme categories to letters on a random basis. The total number of errors for each grade and consonant position was multiplied by each proportion, thus providing an estimate of the number of errors expected to fall into each phonetic distance category under the assumption of randomness. These expected frequencies were statistically compared with the obtained frequencies using the χ^2 analysis. Table 4 presents the obtained and expected frequencies and the value of χ^2 by grade and consonant position.[4]

---

[4]We are aware that the analyses presented in Tables 4 and 5 below, violate the independence assumption of the χ^2 analysis. Consequently, we cannot draw our conclusions from the results of the analysis with any certainty. However, we know of no more appropriate analysis.

TABLE 4: Observed and expected frequencies of consonant errors sharing zero, one, or two features with the target sounds.

Grade	Number of shared features							
	0		1		2			
	Observed	Expected	Observed	Expected	Observed	Expected	χ^2	p
2	11	44	43	65	81	28	132.5	<.001
3	8	25.1	22	37.2	47	16.8	72.2	<.001
4	3	12.4	13	19.6	26	10.5	32.2	<.001

Our expectation that the child's errors would be governed by phonetic features appears to be strongly supported by the consonant data. As can be seen in Table 4, the χ^2 values for each grade and consonant position are significant, with $p < .001$.

The proportion of consonant errors falling into the two-feature-shared category is remarkably stable across the grades: 60 percent of second-grade errors, 61 percent of third-grade errors, and 62 percent of fourth-grade errors share two features with their appropriate target phonemes. The results suggest, therefore, that phonetically motivated substitutions contribute substantially to the consonant error pattern both at the very early stages of reading acquisition and beyond.

FEATURE SUBSTITUTION ERRORS AMONG VOWELS

Vowel errors were treated in much the same way as the consonant errors. A number of feature systems for vowels has been proposed, but none has won such strong empirical support as to give a clear basis for choice. The feature system we used was a modification of that proposed by Singh and Woods (1971). Three of their features--tenseness, tongue advancement, and tongue height--distinguish each nondiphthongized vowel from every other. A fourth feature, retroflexion, distinguishes only the vowel /ɝ/ from other vowels. Since /ɝ/ is an infrequent response in our data, we did not incorporate this feature in our analysis. In its place, we added the feature diphthongization, in order to distinguish diphthongized from nondiphthongized vowels.

The vowel errors, like the consonant errors, were classified according to the number of features they shared with their respective target phonemes. The frequency of errors in each phonetic distance category (zero, one, two, or three features shared with the target) was again compared with the frequencies that would be expected if the child were randomly assigning phonemes to spellings.

The results of the vowel feature analysis, shown in Table 5, reveal a picture very different from the comparable analysis of consonant errors. The Table gives grouped frequencies of errors on the vowel classified according to the number of features shared with the target vowel. Again, expected frequencies are calculated on the null assumption that the distribution of errors within these categories is random. Whereas for consonants the effect of phonetic distance was significant across all grades, the vowel errors displayed in Table 5 reveal no consistent direction in the differences between observed and

expected frequencies. Thus, for vowels, it appears that given the occurrence of an error, the assignment of phoneme to grapheme was random.

TABLE 5: Observed and expected frequencies of vowel errors sharing zero or one feature, or sharing two or three features with the target sounds.

	Number of shared features					
Grade	0-1		2-3			
	Observed	Expected	Observed	Expected	χ^2	p
2	285	304	262	243	2.67	>.10
3	180	191	170	158	1.54	>.20
4	110	118	105	96	1.38	>.20

The contrasting results obtained for vowels and consonants is indeed striking. The opposition of these phonetic classes is revealed by both approaches to error analysis: the first, in which we investigated misreadings in relation to their location in the syllable, and the second, in which we consider the phonetic characteristics of the errors. From the latter analysis, we are led to conclude that the concept of degree of phonetic contrast, so successful in rationalizing the errors on consonants, does not enable us to understand the vowel errors. For these, other sources of difficulty must be sought.

At all events, these differences in error pattern between the consonants and vowels lend credibility to the position taken by ourselves and other investigators (Liberman, Shankweiler, Orlando, Harris, and Bell-Berti, 1971; Vellutino, Steger, and Kandel, 1972; Vellutino, Pruzak, Steger, and Meshoulam, 1973) that visual factors are not sufficient to account for the difficulties of the beginning reader. Surely, problems in scanning, eye movements, and/or the apprehension of the optical form of letters cannot explain the differences in consonant and vowel error patterns that we have found. Consonants and vowels cannot be meaningfully classified in terms of their visual characteristics; the differences in error pattern therefore could not be related to a classification made on that basis. Consonants and vowels do, on the other hand, form distinctive categories in the language and have different functional roles in communication.

Considered from the standpoint of their contribution to the phonological message, consonants carry the heavier information load. Vowels, on the other hand, are the foundation on which the syllable is constructed, and as such are the carriers of prosodic features. It is the vowels that are the more fluid and variable of the two classes of phonetic elements, more subject to phonetic variation across individuals and dialect groups, and more subject to phonetic drift over time. As we suggested in an earlier paper (Shankweiler and Liberman, 1972), the relatively greater variability of vowels than consonants may in part account for the different ways these segments are represented in the orthography. It may account for the fact that in English, at least, there tend to be many spellings for each vowel and more nearly one-to-one spelling-to-sound relationships for the consonants.

SUMMARY AND CONCLUSIONS

The errors children make in reading, before they have fully mastered the skill, can teach us something about the special problems of learning to read. In an earlier study, we observed, as others have, that errors on the final consonant of a CVC syllable far exceed those on the initial consonant. Additionally, we found that errors on medial vowels exceed those on consonants in both initial and final position. The first purpose of the present study was to confirm these earlier findings and, by the use of various controls, to test their generality.

We found the same pattern of consonant errors as previously obtained, with those in final position being misread twice as often as those in initial position. As a result of the controls introduced in the present study, we can now conclude that the findings represent a true position effect. It cannot be attributed to a different phonological distribution of consonants in syllable-initial and in syllable-final position, nor can it be attributed to differences in the orthography associated with beginnings and ends of words. Having ruled out these interpretations of the position effect, we believe the greater difficulty of the final consonant is the result of the child's defective understanding of the phonological segmentation of his spoken language. We know from earlier work of our own and others that inability to indicate the phonemic segmentation of heard speech is characteristic of the prereading child. Given the difficulty in becoming explicitly aware that syllables may be analyzed into strings of phonological segments, it seemed reasonable to suppose that the task of phonological segmentation might vary in difficulty with the position of a given segment in the syllable. On this basis, the initial segment should be easiest to isolate because it can be extracted without analysis of the internal structure of the syllable.

In contrast to the findings on consonant misreadings, errors on vowels show no effect of position. When we placed the vowel in initial, medial, and final position in the syllable, the errors did not vary in any systematic fashion. We suppose, therefore, that vowel errors do not reflect primarily the child's difficulties in phonological segmentation, but rather the complexity and variability of the spelling-to-sound correspondences.

The assumption that consonant and vowel errors have different causes was supported by the results of a further analysis that took account not of the location of the errors, but of their nature. In that analysis, it was found that consonant errors were systematically related to the presented phoneme, differing from it most often in only one feature. Vowel errors, in contrast, were not systematically related to the phonetic features of the presented vowel; indeed, the feature distribution of the vowel errors was essentially random. Such differences in the distribution of errors on consonants and vowels in reading may reflect the different functions of those phonetic classes in speech. Perhaps the most general implication of these differences in error pattern between consonants and vowels is that they underscore the importance of nonvisual cognitive processes in reading. These findings lend confirmation to our belief that visual factors contribute rather little to the difficulties of beginning reading--certainly less than factors relating to the language, such as awareness of phonological segmentation, phonetic recoding, and the structure of the orthography.

REFERENCES

Buckingham, B. and E. Dolch. (1936) A Combined Word List. (Boston: Ginn & Co.).

Calfee, R., R. Chapman, and R. Venezky. (1972) How a child needs to think to learn to read. In Cognition in Learning and Memory, ed. by L. W. Gregg. (New York: Wiley).

Daniels, J. C. and H. Diack. (1956) Progress in Reading. (Nottingham, England: University of Nottingham).

Eimas, P. D. (in press) Distinctive-feature codes in the short-term memory of children. J. Exp. Child Psychol.

Gibson, E. and H. Levin. (1975) The Psychology of Reading. (Cambridge, Mass.: MIT Press).

Liberman, I. (1971) Basic research in speech and lateralization of language: Some implications for reading disability. Bull. Orton Soc. 21, 71-87.

Liberman, I. (1973) Segmentation of the spoken word and reading acquisition. Bull. Orton Soc. 23, 65-77.

Liberman, I., D. Shankweiler, F. W. Fischer, and B. Carter. (1974) Reading and the awareness of linguistic segments. J. Exp. Child Psychol. 18, 201-212.

Liberman, I., D. Shankweiler, C. Orlando, K. Harris, and F. Bell-Berti. (1971) Letter confusions and reversals of sequence in the beginning reader: Implications for Orton's theory of developmental dyslexia. Cortex 7, 127-142.

Miller, G. and P. Nicely. (1955) An analysis of perceptual confusions among the English consonants. J. Acoust. Soc. Am. 27, 338-352.

Rosner, J. and D. P. Simon. (1970) The Auditory Analysis Test: An Initial Report. (Pittsburgh: University of Pittsburgh Learning Research and Development Center).

Rozin, P. and L. Gleitman. (in press) The structure and acquisition of reading. In Reading: Theory and Practice, ed. by A. S. Reber and D. Scarborough. (Hillsdale, N. J.: Lawrence Erlbaum Assoc.).

Shankweiler, D. and I. Liberman. (1972) Misreading: A search for causes. In Language by Ear and by Eye: The Relationships between Speech and Reading, ed. by J. F. Kavanagh and I. G. Mattingly. (Cambridge, Mass.: MIT Press).

Singh, S. and D. R. Woods. (1971) Perceptual structure of 12 American English vowels. J. Acoust. Soc. Am. 49, 1861-1866.

Vellutino, F., R. Pruzak, J. Steger, and U. Meshoulam. (1973) Immediate visual recall in poor and normal readers as a function of orthographic-linguistic familiarity. Cortex 8, 106-118.

Vellutino, F., J. Steger, and G. Kandel. (1972) Reading disability: An investigation of the perceptual deficit hypothesis. Cortex 9, 370-386.

Weber, R. (1970) A linguistic analysis of first-grade reading errors. Read. Res. Quart. 5, 427-451.

Comments on the Session: Perception and Production of Speech II; Conference on Origin and Evolution of Language and Speech*

A. M. Liberman+

The interesting papers we heard all dealt in one way or another with a question that is surely central to an inquiry into the biology of language: Are linguistic processes in some sense special, different from the processes that underlie nonlinguistic activities and, perhaps, unique to man? To discuss that question, and the papers of the evening's session, I find it useful to distinguish two classes of specialized processes, auditory and phonetic.

Specialized auditory processes would serve, perhaps in the fashion of feature detectors, to extract those aspects of the acoustic signal that carry the important information. One is led to suppose that such devices might exist because it is true, and paradoxical, that some of the most important phonetic information is contained in parts of the speech sound that are not physically salient. Thus, a significant acoustic cue is in the formant transitions, though these are often of short duration and rapidly changing frequency. Perhaps there are devices devoted to detecting those transitions. If so, we should hold them up as examples of specializations in the auditory system. They would be important for the perception of language, but not properly part of its special processes.

If the acoustic signal were directly related to the phonetic message, then detection of the phonetically important cues would be sufficient for phonetic perception; no further processing would be necessary. But the relation between signal and message is peculiarly complex. [For summary accounts, see Fant (1962); Cooper (1972); Stevens and House (1972); Liberman (1974); Studdert-Kennedy (1974).] As a result, the specialized auditory detectors can only begin the job; the auditory display they produce must still be interpreted, because the phonetic message is there in such highly encoded form. If there are devices specialized to do that kind of interpreting, then I should consider them phonetic, not auditory. Since I will organize my comments on the papers of the evening in terms of that distinction, I should take a moment to illustrate what I mean.

Consider the formant transitions that are important cues for the perception of stop consonants in syllable-initial position, and call up in your mind's eye spectrographic representations (similar to those shown by Dr. Morse) of such transition cues as would be appropriate for [da] and [ba]. Now add a patch of

---

*Paper delivered at the New York Academy of Sciences, 22-25 September 1975.

+Also University of Connecticut, Storrs, and Yale University, New Haven, Conn.

[HASKINS LABORATORIES: Status Report on Speech Research SR-45/46 (1976)]

fricative noise--the hiss of [s]--just before the [da]. If that patch is immediately in front of the [da], you will hear [sa], not [da]; the stop will have disappeared completely. But if the patch is moved away so as to leave about 50 msec of silence between the end of the hiss and the beginning of the formant transitions, then you will hear [sta]; that is, you will hear the stop once again. The generalization that captures those facts, and many others closely related to them, is that a necessary condition for the perception of syllable-initial stop consonants is a brief period of silence in front of the appropriate transition cues. But why should silence be necessary? Why should it be impossible to hear the stop when its acoustic cues follow closely on the fricative noise?

The simplest explanation, surely, is that we are here dealing with a characteristic of the generalized mammalian auditory system. That might seem reasonable if only because in putting the fricative noise in front of the transition cues we have conformed to the paradigm for auditory forward masking. But a search of the literature on such masking uncovers no reason to suppose that it could, in fact, provide the account we seek; forward masking does occur, but it is not nearly so strong as to produce the total disappearance of the stop consonant in [sa]. [See, for example, Elliott (1971) and Leshowitz and Cudahy (1973).]

Consider, now, a second interpretation. Suppose there are transition detectors of the kind I speculated about and suppose, further, that the fricative noise disables them, rendering them ineffective in extracting the transition cues for the stop consonant. In fact, there is very indirect evidence that such transition detectors may exist in man. Thus, work by Kay and Matthews (1972) suggests that there may be detectors sensitive to frequency modulations, at least within a certain range. More, and perhaps more indirect, evidence comes from studies on the so-called adaptation-shift phenomenon, first found in speech by Eimas and Corbit (1973) and since studied by a number of investigators. [For a review, see Cooper (1975) and Darwin (in press).] Among those studies is a recent one by Ganong (1975) that I will describe, if only briefly, because its outcome has several implications for our concern with specialized processes: it suggests, as do several other such studies, that transition detectors may exist, but it also indicates that such detectors are in no way disabled by the fricative noise of our example.

Ganong's experiment went like this. Having first found the boundary between synthetic [da] and [ba], Ganong adapted his subjects with [da] and measured the resulting shift in the [da-ba] boundary. Then he put a patch of fricative noise in front of the [da] and adapted his subjects with the [sa] syllable that they all heard when the fricative-patch-plus [da] was sounded. The effect on the [da-ba] boundary was at least as great as when the adaptation was carried out with [da]. As a control against the possibility that [sa] had its effect because it worked on the same abstract phonetic-feature detector as [da] ([s] and [d] have the same place-of-production feature), Ganong adapted with a [sa] from which the formant transitions had been removed; in that condition the effect on the [da-ba] boundary was much smaller. Those results suggest that the adaptation shift in the [da-ba] boundary was caused by a change in the state of some device that responds to formant transitions; thus, they support the assumption that there are such things as transition detectors.

But Ganong's results also show, more generally, that the transition cues following the fricative noise were getting through in full strength, at least as auditory events. If those transition cues nevertheless failed to produce perception of a stop consonant, it was not because they were absent from the auditory display. [Other kinds of evidence for the same conclusion are reviewed in Liberman (in press).]

We are led, then, to a third explanation for the disappearance of the stop consonant: silence is necessary for the perception of stop consonants, not because it provides time to evade normal auditory forward masking, and not because it prevents the disabling of specialized transition detectors, but because it provides information. The information is that the speaker did indeed make the total closure of the vocal tract necessary to the production of a stop consonant. Thus, given enough silence to indicate a sufficient closure of the vocal tract, a specialized phonetic device could interpret the transition cues as reflecting a linguistic event that included the stop-consonant segment [d]. Hence the perception [sta] when a silent interval of about 50 msec is placed between the end of the hiss and the beginning of the transitions. Without that silent interval the only reasonable phonetic interpretation is that the vocal tract did not close completely. Hence [sa].

So much, then, for the possibility that there are at least two different kinds of devices specialized for speech. Let me now comment on the papers of the evening with reference to that distinction.

In the presentation by Dr. Andrews we saw interesting evidence that baboons change the configuration of their vocal tracts so as to produce something like formant transitions and, further, that such transitions may convey information from one baboon to another. If it is indeed the formant transitions that carry the information, and if the transitions are as brief and rapid as they sometimes are in human speech, then we should not be surprised to find feature detectors specialized to tract them. And in working with baboons we might, of course, expect to get at such devices more directly than we can in research on human beings.

Though baboons may produce and respond to rapid transitions, we have as yet found no reason to believe that they (or, indeed, any creatures other than man) produce or perceive phonetic strings. I should doubt, therefore, that we would find the specialized phonetic processor to which I referred. But what I doubt is surely not important. What is important, I should think, is that we can find out whether baboons do have something like transition detectors and also whether they behave toward speech as if they make a phonetic interpretation. Dr. Andrews has given us a good start in that direction.

The experiments that Philip Morse described are a model of how to learn about the biology of language. To select some interesting characteristic of human speech perception and then look for that characteristic in prelinguistic infants and nonhuman primates is surely one of the best ways to uncover whatever there may be of biological predisposition, specialized process, and species specificity. The experiments are certainly hard to do, but they are very much worth doing, and Dr. Morse does them very well indeed.

The results Dr. Morse told us about this evening were interpreted by him in terms of the possibility that there are devices like transition detectors. In

his view, such devices might explain categorical perception of the place distinction for stop consonants in infants and the somewhat in-between tendency toward categorical perception he got in monkeys. I think it quite reasonable to suppose that the output of such detectors would be categorical. I doubt, however, that the concept of feature detector could take us very far toward explaining the perception of stop consonants, except by a kind of metaphorical extension. Some of the reasons for my doubt will, perhaps, become clearer in connection with the examples I mean to develop when I discuss Dr. Warren's paper in a few moments, so I will say no more about those reasons now. In fairness to Dr. Morse, however, I should emphasize that he was not trying to explain the perception of stop consonants, nor even the perception of the place feature, but only some data on discrimination and tendencies toward categorical perception in infants and monkeys.

As for Dr. Morse's experiment, I should say that in using three formants instead of two he gained the advantage of greater realism but at the cost of some added difficulty in interpretation of the results. That difficulty arises because when second- and third-formant transitions are both varied, it is harder to scale physical similarity and therefore that much harder to assess tendencies toward categorical perception. If one nevertheless prefers to use the three-formant patterns because they are closer to what occurs in speech, he might reduce the difficulty I referred to by coupling the transition cues with a variety of vowels, thus randomizing the acoustical similarities; if the discrimination functions nevertheless come out the same way they did in Dr. Morse's experiment, the conclusion would be quite compelling.

Still, the results so far obtained with infants are impressive. The infants of Morse's study did show a strong tendency toward categorical perception of the place distinction in the stops, and, as Morse pointed out, that result accords with those obtained by other investigators. In the case of the monkeys, however, it is a good deal less clear that perception of the stops is categorical. There was, in the monkeys of Dr. Morse's experiment, some tendency in that direction, though less apparently than with the infants. In that connection, we should keep in mind the results of the earlier study by Sinnott (1974), to which both Morse and Warren referred. Using reaction time as the measure, Sinnott found that her monkeys, like those of Morse, discriminated within phonetic categories; but they did not discriminate better across phonetic boundaries than within them. That is, Sinnott's monkeys did not show any appreciable tendency toward categorical perception, though her human subjects did.

Since the experiment on discrimination of the voicing distinction by chinchillas (Kuhl and Miller, 1975) was several times referred to by our speakers, I should also comment on that. It is surely of interest that the chinchillas "classified" the speech stimuli so as to put the boundary in much the same place that human listeners do. Given that the relevant acoustic cue is the relative time of onset of two parts of the pattern, it is also of interest that research with nonspeech sounds has found a categorical "notch" in the auditory system at a relative displacement appropriate to the speech-sound boundary (Miller, Pastore, Wier, Kelly, and Dooling, 1974). In the case of the voicing distinction, it may be, therefore, that in the development of language, nature took advantage of a categorical distinction characteristic of some mammalian auditory systems, though special adjustments in the articulatory mechanisms would presumably have been necessary to get them to produce accurately just that

small difference in timing required to put the sounds within the preset (and rather narrow) constraints of the ear.

I nevertheless have several reservations, even about this apparently simple case. Using an expanded range of the same stimuli that were used in the chinchilla experiment, Wilson and Waters (1975) found that variations in stimulus range caused rhesus macaque monkeys to shift their "boundary" from 28 msec, which happens to be about where the chinchilla boundary was, to 66 msec. (They also found some tendency toward categorical perception, wherever the boundary was.) That kind of change, which implies that the monkeys may have been splitting the range, does not occur in human subjects. [See, for example, Sawusch, Pisoni, and Cutting (1974).] The possibility that such a change might occur in chinchillas was not controlled for.

My other reservation arises from the fact that the human boundary is not fixed at either of the boundaries so far found with animals and with nonspeech sounds, but rather varies (together with the categorical notch) from 18 msec to as much as 45 msec as a function of the duration of the transitions and the frequency at which the first formant begins (Stevens and Klatt, 1974; Lisker, Liberman, Erickson, and Dechovitz, 1975). (The variation with duration of the transitions may reflect a normalization for rate of articulation.) I would be interested to know if the chinchilla's boundary moves in the same way. It would also be interesting to know if the chinchilla, or any other animal, appreciates that the voicing distinction is, indeed, the same in those cases in which the relevant acoustic cues are entirely different. What happens, for example, when the distinction is moved from initial position (e.g., [bɪ] vs. [pɪ], which is the kind of distinction so far studied in animals) to intervocalic position following a stressed syllable (e.g., [ræbɪd] vs. [ræpɪd]), where a sufficient cue is the time interval between the two syllables; or to final position [e.g., [raeb] vs. [raep], where a sufficient cue is the duration of the preceding vowel (plus consonant-vowel transition)]? To "understand" that such distinctions have something in common despite gross difference in the acoustic cues would constitute an impressive demonstration of phonetic interpretation.

We come now to that part of this evening's program that touched more directly on the matter of specialized phonetic processes. The relevant paper was given by Richard Warren. He reminded us of his earlier experiments--very important experiments, in my view--in which he found that the auditory system does not measure up to one of the requirements of phonetic perception. The requirement is that the order of the phonetic segments be preserved; the word "bad" is different from the word "dab." Now if we measure the rates at which speech is produced and perceived, we find that the durations we can allot to the phonetic segments are often very short. Indeed, those durations can be as little as 50 msec per segment or, for brief periods, even less. But Dr. Warren has found with nonspeech sounds that the ear cannot properly cope with segments of those temporal dimensions. At the very short durations that we can assign to phonetic segments, the ear can discriminate one order of segments from another--that is, it can hear distinctively different patterns--but, as Dr. Warren told us, it is unable to identify the separate components in the order of their occurrence. Now I will not here review or comment on Dr. Warren's solution to this very real problem. I will rather offer an alternative, which is that in perceiving the order of the phonetic segments we need not--and indeed do not--rely on the temporal order of acoustic segments. Indeed, I would argue that even if the ear were able to identify the order of very short-duration acoustic

segments, it could hardly make use of that ability in perceiving speech. That would be so because the string of phonetic segments is drastically restructured in the conversion to sound, with the result that segmentation of the sound does not correspond directly to the segmentation of the message; accordingly, the segments are not signaled simply by acoustic events in ordered sequence. But, fortunately for the integrity of the message, information about segment order is nevertheless conveyed, though by acoustic cues that could be interpreted, I should think, only by a device that "knows" the secret of the code--that is, by a phonetic device.

Let us consider, for example, the matter of segment order in the syllables [ba] and [ab] and see how information about the phonetic structure is carried in the sound. In producing those syllables, the gestures for the segments [b] and [a] are not made discretely and in turn. Rather, as we well know, the gestures are organized into units larger than a segment--something like a syllable, perhaps--and then coarticulated. If the [ba] and [ab] syllables had been produced at a moderately high rate of articulation, we should then see for [ba] an acoustic signal lasting perhaps 70 or 80 msec and containing three formants that rise from the beginning of the acoustic syllable to the end. For [ab] we should see the mirror image--that is, three formants that fall. If we search out the information about [b], we find that it exists not just at the beginning (for [ba]) or at the end (for [ab]), but throughout the acoustic syllable. Information about the vowel is also carried from one end of the sound to the other. It is as if the coarticulation has effectively folded consonant and vowel into the same piece of sound. As a result, there is no acoustic criterion by which one can divide the speech signal into segments corresponding to the segments of the phonetic message. A further consequence is that the cues for the segments must necessarily exhibit a great deal of context-conditioned variation: the transition cues for the consonant, for example, are rising in the one case and falling in the other. (It should be remarked that when we listen to those transitions in isolation we hear rising and falling glissandos, just as our knowledge of auditory psychophysics would lead us to expect.)

To explain how a listener might recover the identity of the segments--that is, know that there is a consonant [b] and a vowel [a]--we might suppose that there is a specialized phonetic device that can "hear through" the context-conditioned variation in the acoustic cues and arrive at the canonical forms of the segments. If so, then that same device could use the same context-conditioned variation to discover the order of the segments: for if the rising pattern contains a [b], then it could only be a syllable-initial [b]; and if the falling pattern contains a [b], it could only be a syllable-final [b]. Thus, I would suppose that perceiving the order of the phonetic segments does not depend on the ability of the ear to deal with discrete sounds of short duration, but rather on the operation of a special phonetic device that is able to cope with the fact that information about order is often encoded in the sound as variations in acoustic shape. Indeed, I would suppose that such encoding would seem nicely designed to evade just those limitations of the ear that Dr. Warren's research has revealed.

I should comment finally on the paper by Philip Lieberman. His work is especially interesting from my point of view because it offers evidence for a specialization associated with the production of speech that is, in an important sense, analogous to the transition detectors of the auditory system. To see the analogy, we should consider what might have occurred as grammar--hence language--

evolved. The view I want to present has been developed elsewhere (Liberman, 1974), so I will only outline it here.

If, as in an agrammatic system of acoustic communication, the messages were directly linked to sounds, the number of messages we could communicate would be limited to the number of holistically different sounds we can produce and perceive. And that is a relatively small number. But grammar drastically restructures the information in the message, making it appropriate, at the one end, for the great message-generating capabilities of the brain and, at the other, for the relatively limited abilities of the vocal tract and the ear to produce and perceive sounds. Viewed this way, the processes underlying grammar evolved as a kind of interface between two different kinds of structures, adapting the potentialities of the one to the limitations of the other. (My earlier comments on evading the auditory limitations described by Dr. Warren are an example of this kind of grammatical function at the very lowest level of the linguistic system--that is, at the conversion from phonetic message to sound.) But it is also possible that in this evolutionary process the structures being linked by the grammar might themselves have changed. On the perception side of the process an example would be the development of transition detectors in the auditory system to extract just that information which the phonetic (grammatical) system used in carrying out its peculiar function. And on the production side there are the changes in the vocal tract that Dr. Lieberman has told us about. Those changes have apparently made the vocal tract less limited for phonetic communication, and so have reduced the mismatch between that organ and the message-generating intellect, a mismatch otherwise taken care of by the grammar. We might suppose that if we had to speak with the vocal tract of a nonhuman primate, the grammatical interface would have to be even more complex than it is.

I think I can justifiably end my comments on a hopeful note. Those of us who care about speech and the biology of language have reason to be encouraged. We now know enough about speech to be able to identify some of its most distinctive characteristics--those characteristics, that is, that most clearly imply the existence of specialized linguistic processes. As a result, we can fruitfully make comparisons with nonlinguistic processes in man and with any processes at all in prelinguistic infants and (presumably) nonlinguistic animals. Indeed, the comparisons are, for obvious reasons, easier to make at the level of speech than at the level of syntax, especially with infants and animals. Moreover, we have started to make those comparisons. But we have only just started. There are hundreds of experiments out there waiting to be done. Until we see what results they produce, we would be well advised, I think, to suspend judgment.

REFERENCES

Cooper, F. S. (1972) How is language conveyed by speech? In Language by Ear and by Eye, ed. by J. F. Kavanagh and I. G. Mattingly. (Cambridge, Mass.: MIT Press), pp. 25-45.

Cooper, W. E. (1975) Selective adaptation to speech. In Cognitive Theory, vol. 1, ed. by F. Restle, R. M. Shiffrin, N. J. Castellan, H. R. Lindman, and D. B. Pisoni. (Hillsdale, N. J.: Lawrence Erlbaum Assoc.), pp. 23-54.

Darwin, C. J. (in press) The perception of speech. In Handbook of Perception, vol. 7, ed. by E. C. Carterette and M. P. Friedman. (New York: Academic Press). [Also in Haskins Laboratories Status Report on Speech Research SR-42/43 (1975), 59-102.]

Eimas, P. D. and J. D. Corbit. (1973) Selective adaptation of linguistic feature detectors. Cog. Psychol. 4, 99-109.

Elliott, L. L. (1971) Backward and forward masking. Audiology 10, 65-76.

Fant, C. G. M. (1962) Descriptive analysis of the acoustic aspects of speech. Logos 5, 3-17.

Ganong, W. F. (1975) An experiment on "phonetic adaptation." Quarterly Progress Report (Research Laboratory of Electronics, MIT) 116, 206-210.

Kay, R. H. and D. R. Matthews. (1972) On the existence in human auditory pathways of channels selectively tuned to the modulation present in frequency-modulated tones. J. Physiol. (London) 225, 657-677.

Kuhl, P. K. and J. D. Miller. (1975) Speech perception by the chinchilla: Voiced-voiceless distinction in alveolar plosive consonants. Science 190, 69-72.

Leshowitz, B. and E. Cudahy. (1973) Frequency discrimination in the presence of another tone. J. Acoust. Soc. Am. 54, 882-887.

Liberman, A. M. (1974) The specialization of the language hemisphere. In The Neurosciences: Third Study Program, ed. by F. O. Schmitt and F. G. Worden. (Cambridge, Mass.: MIT Press), pp. 43-56. [Also in Haskins Laboratories Status Report on Speech Research SR-31/32 (1972), 1-22.]

Liberman, A. M. (in press) How abstract must a motor theory of speech be? Paper delivered at the 8th International Congress of Phonetic Sciences, Leeds, 21 August 1975. [Also in Haskins Laboratories Status Report on Speech Research SR-44 (1975), 1-15.]

Lisker, L., A. M. Liberman, D. Dechovitz, and D. Erickson. (1975) On pushing the voice-onset-time boundary about. J. Acoust. Soc. Am., Suppl. 57, S50(A). [Also in Haskins Laboratories Status Report on Speech Research SR-42/43 (1975), 257-264.]

Miller, J. D., R. E. Pastore, C. C. Wier, W. J. Kelly, and R. J. Dooling. (1974) Discrimination and labeling of noise-buzz sequences with various noise-lead times. J. Acoust. Soc. Am., Suppl. 55, S390(A).

Sawusch, J. R., D. B. Pisoni, and J. E. Cutting. (1974) Category boundaries for linguistic and nonlinguistic dimensions of the same stimuli. Research on Speech Perception (Department of Psychology, Indiana University) 1, 162-173.

Sinnott, J. M. (1974) Human versus monkey discrimination of the /ba/ /da/ continuum using three-step paired comparisons. J. Acoust. Soc. Am., Suppl. 55, S55(A).

Stevens, K. N. and A. S. House. (1972) The perception of speech. In Foundations of Modern Auditory Theory, vol. 2, ed. by J. Tobias. (New York: Academic Press), pp. 3-62.

Stevens, K. N. and D. H. Klatt. (1974) Role of formant transitions in the voiced-voiceless distinction for stops. J. Acoust. Soc. Am. 55, 653-659.

Studdert-Kennedy, M. (1974) The perception of speech. In Current Trends in Linguistics, ed. by T. A. Sebeok. (The Hague: Mouton). [Also in Haskins Laboratories Status Report on Speech Research SR-23 (1970), 15-48.]

Wilson, W. A. and R. S. Waters. (1975) How monkeys perceive some sounds of human speech. Paper read at a meeting of the American Psychological Association, Chicago, September.

Consonant Environment Specifies Vowel Identity*

Winifred Strange,[+] Robert R. Verbrugge,[+] Donald P. Shankweiler,[++] and Thomas R. Edman[+]

ABSTRACT

Past studies have shown that while vowels can be produced with static vocal-tract configurations, the resulting steady-state tokens are misidentified frequently by naive listeners. The first experiment compared the perception of isolated vowels with vowels spoken in a fixed consonantal frame by the same set of 15 talkers. Vowels in /p-p/ syllables were identified with far greater accuracy than were comparable isolated vowels in both single and multiple talker conditions. Acoustical analyses of the test tokens showed that the poor intelligibility of isolated vowels could not be attributed to talkers' failure to produce these vowels correctly. In a second experiment, vowels in syllables in which the initial and final stop consonant varied unpredictably from item to item were still identified with greater accuracy than were isolated vowels. These results offer strong evidence that dynamic acoustic information distributed over the temporal course of the syllable is used regularly by the listener to identify vowels.

---

*A partial summary of these results was presented at the 87th meeting of the Acoustical Society of America, New York, 25 April 1974, and published in Strange, Verbrugge, and Shankweiler (1974). A more complete exposition of the problem of perceptual constancy in speech perception may be found in Shankweiler, Strange, and Verbrugge (in press).

[+]University of Minnesota, Minneapolis.

[++]Also University of Connecticut, Storrs.

Acknowledgment: This paper reports research begun during the academic year 1972-1973 while D. Shankweiler was a guest investigator at the Center for Research in Human Learning, University of Minnesota, Minneapolis. The work was supported by grants to the Center and to Haskins Laboratories from the National Institute of Child Health and Human Development, by grants awarded to D. Shankweiler and J. J. Jenkins by the National Institute of Mental Health, and by a fellowship to R. Verbrugge from the University of Michigan Society of Fellows. We wish to thank Kevin Jones, Kathleen Briggs, and Robert Jenkins for their assistance in the experimental work, and James Jenkins for his advice and encouragement throughout this research.

[HASKINS LABORATORIES: Status Report on Speech Research SR-45/46 (1976)]

INTRODUCTION

Vowels, unlike consonants, can be produced and identified in isolation. This possibility was exploited early in the investigation of vowel quality, as witnessed by studies of the cardinal vowels (Jones, 1956). Sustained, "steady-state" vowels can be classified by frequencies of the first two or three formants (Potter and Steinberg, 1950). So successful were the efforts to locate the acoustic information sufficient for the perception of sustained vowels that the main focus of research on speech perception shifted to the search for the consonantal cues. But the supposition that the sound pattern is simpler in the case of the vowels than the consonants is unsupportable if a distinction is made between the sustained, isolated vowel and the vowel as it occurs in natural speech.

Although they can be produced in a quasi-steady-state manner and in isolation, vowels so produced must be regarded as laboratory artifacts. Ordinarily, vowels occur in coarticulation with consonants in the context of the syllable. The acoustic information in coarticulated vowels is fused and carried in parallel with the consonantal information. (See Liberman, Cooper, Shankweiler, and Studdert-Kennedy, 1967; Liberman, 1970.) It was discovered long ago in tape-cutting experiments of Schatz (1954) and Harris (1953) that vowel quality cannot be discretely localized in any single portion of the syllable, but is distributed throughout the period during which voicing is present.

Studies of perturbations of formant frequencies brought about by uttering vowels in the context of syllables were carried out by Shearme and Holmes (1962), Lindblom (1963), Stevens and House (1963), and Öhman (1966). These investigations demonstrated that steady-state values of the formants are rarely attained because articulatory movement is more or less continuous. Thus, the acoustic description of vowels in ordinary speech is a good deal more complex and problematic than is revealed by the classic studies of the acoustic basis of vowel quality.

If the acoustic structure of the isolated vowel often differs greatly from the "same vowel" in context, it might be inferred that different cues are employed in vowel perception when the vowel is in consonantal context and when it occurs in isolation. It is all the more interesting, therefore, to find indications in the phonetic literature that isolated vowels are difficult to perceive. For example, Fairbanks and Grubb (1961) presented nine isolated vowels produced by phonetically trained talkers to experienced listeners. The overall identification rate was only 74 percent, which contrasts strikingly with a rate of 94 percent obtained by Peterson and Barney (1952) for perception of vowels in /h-d/ context. Somewhat better identification of isolated vowels was obtained by Lehiste and Meltzer (1973), with only three talkers producing the tokens. Fujimura and Ochiai (1963) directly compared the identifiability of vowels in consonantal context and in isolation. They found that the center portions of vowels, which had been gated out of consonant-vowel-consonant (CVC) syllables, were less intelligible in isolation than in syllabic context. These findings suggest that isolated vowels are misidentified with significantly higher frequency than vowels spoken in at least some consonantal environments. Could it be that the acoustic complexities introduced by syllabic structure better serve the requirements of the perceptual apparatus than do quasi-steady-state formants? If so, then it is surely inappropriate to characterize the cues for vowel identity in terms of static points in a space defined by the first two formants.

It seemed important, therefore, to attempt to demonstrate under carefully controlled experimental conditions that vowels in consonantal contexts are perceived with fewer errors than "the same vowels" presented in isolation. A further purpose of the research reported here was to investigate the sources of information within the CVC syllable that specify the vowel and to explore how that information is used by the perceiver in the process of perception. If it is true that consonantal environment generally aids in identification of a vowel, we recognize that there is more than one way the environment might play a facilitating role. One possibility is that portions of the signal commonly regarded as consonantal, such as transitions, might aid in normalization for vocal-tract differences. Experiments by Fourcin (1968) and Rand (1971) have found that perceptual boundaries between stop consonants vary depending on the vocal tract presumed to have produced a syllable. The phonemic identity of the consonants was fixed and known in advance in the Peterson and Barney (1952) study and in our own investigations (Verbrugge, Strange, Shankweiler, and Edman, in press). In these cases, the transitions may have allowed listeners to scale the formant frequencies of the medial vowel according to the vocal-tract characteristics of the talker and thus reduce vowel ambiguity.

On the other hand, isolated vowels may be difficult to perceive for a more fundamental reason. It is possible that listeners ordinarily rely upon information distributed throughout the whole syllable for identification of the vowel. This seems likely in view of parallel transmission of the consonants and the vowel. If it is the case that syllable-initial and syllable-final transitions specify the vowel as well as the consonants, we could assert that the vowel is inseparable from the syllable, that it is not specified by formant frequencies at any particular cross section in time, but rather is carried in the dynamic configuration of the whole syllable. In this case the presence of transitions should aid identification of the intended vowel whatever additional difficulties may be posed by confronting the listener with multiple vocal tracts.

EXPERIMENT I: PERCEPTION OF ISOLATED AND MEDIAL VOWELS

If consonantal environment aids in specifying vowel identity in either of the two ways postulated above, we would expect that the perception of isolated vowels would be less accurate than the perception of medial vowels in listening tests where the tokens on a test were produced by different talkers. Previous studies on the identification of steady-state vowel stimuli support this hypothesis (Fairbanks and Grubb, 1961; Lehiste and Meltzer, 1973). However, these investigations do not directly compare isolated vowels with vowels in syllable frames when the number and type of talkers, number of response alternatives, and other factors are held constant. Millar and Ainsworth (1972) report that listeners were able to identify synthetically generated vowels more reliably and uniformly when the vowels were embedded in /h-d/ words than when the acoustically identical segments were presented in isolation. We are not aware of any studies that directly compare the perception of naturally produced isolated vowels with vowels in context.

The present study compares the identifiability of vowels produced in a fixed consonantal frame with isolated vowels when (1) a single talker produced all tokens on a particular listening test (Segregated Talker condition) and (2) when tokens produced by several different talkers are presented in random order (Mixed Talker condition). By independently varying these two factors

(consonantal context and talker variation), we can assess the relative contribution of each to the accuracy of vowel identification. Further, the design allows us to test the two hypotheses regarding the way in which consonantal information may be utilized. If consonantal environment aids in vowel identification by serving as a calibration signal for vocal-tract normalization, we expect an interaction between the two major variables. That is, we expect that the loss in identifiability of vowels due to the absence of consonantal transitions will be more severe in those tests where talker identity changes, since recalibration is necessary on each trial. We expect no significant disadvantage of the absence of consonantal transitions for those tests in which talker identity is unchanged. Alternatively, if consonantal transitions provide information that specifies vowel identity independent of talker normalization, we expect no such interaction. The identification of isolated vowels should be less accurate than of vowels in consonantal context both for tests on which the talker remains constant and for tests on which talkers are mixed.

This study compares listeners' performance on isolated vowel tests with the results reported previously for medial vowels spoken in /p-p/ environment (Verbrugge et al., in press). The tests were directly comparable on all factors, such as identity of talkers, order of presentation of alternatives, response alternatives, and recording and reproduction conditions.

Method

Stimulus materials. The panel of talkers described in our previous research was also used for this study. Five men, five women, and five children, none of whom were trained speakers, were selected to represent a wide variety of vocal-tract sizes and characteristic fundamental frequencies. According to the judgment of the experimenters, the talkers represented a fairly homogeneous dialect group, that of the upper midwest region of the United States from which the listeners were also drawn.

The materials for the /p-p/ tests (Mixed and Segregated Talker) were those described in Verbrugge et al. (in press: Exp. II). Talkers read the test syllables, which were printed individually on cards. The /p-p/ words were also used to represent the isolated vowels; talkers were instructed to pronounce the vowels as they would be pronounced in these key words. They were given one practice trial and were instructed to produce the tokens quite rapidly. Each talker produced one token of each of nine isolated vowels: /i/, /ɪ/, /ɛ/, /æ/, /ɑ/, /ɔ/, /ʌ/, /ʊ/, /u/.

For the Mixed Talker Isolated Vowel test (Mixed #-#) three of the nine vowels were selected for each talker, corresponding to the three vowels he produced for the /p-p/ test. As in the earlier test, vowels were assigned to talkers randomly with the constraint that each talker contributed only one of the point vowels. Thus, the Mixed #-# test consisted of five tokens of each of nine vowels; each of the five tokens was spoken by a different talker.

The Segregated Talker Isolated Vowel tests (Segregated #-#) were comparable to the Segregated Talker /p-p/ tests described in Verbrugge et al. (in press: Exp. II). One man, one woman, and one child each produced a 45-item test that contained five different tokens of each of the nine vowels.

All test stimuli were recorded in a sound-attenuated experimental room with a ReVox A77 stereo tape recorder and Spher-o-dyne microphone. The 45 tokens on a test were arranged in a random presentation order with the restrictions that the same intended vowel did not appear more than twice consecutively, and tokens produced by the same talker were separated by not less than eight tokens (in the Mixed tests). Identical procedures were used to construct each of the four tests so that presentation order, timing, and peak intensity of test tokens were identical for all tests.

Procedure. Listening tests were presented to small groups of subjects in a quiet experimental room via a Crown CX 822 tape recorder, MacIntosh MC40 amplifier, and AR acoustic suspension loudspeaker. Listeners responded on score sheets that contained nine response alternatives written out in full in each row: "pip, pup, pap, peep, pop, pep, poop, pawp, puup." Before the tests, the experimenter pronounced each of the nine key words, drawing special attention to the last word, "puup," which stood for the syllable /pʊp/. For the #-# tests, the experimenter pronounced each key word followed by the vowel in isolation, again with special attention to the /ʊ/ alternative. Subjects in the Mixed Talker conditions were told they would hear "several different talkers"; subjects in the Segretated Talker conditions knew thay would hear only one voice on each 45-token test.

Independent groups of subjects responded to the /p-p/ and the #-# Mixed Talker tests. Each group of subjects completed two repetitions of the 45-token test for a total of 90 judgments per subject, 10 on each intended vowel. In the Segregated Talker conditions, three groups of subjects heard the /p-p/ tests and another three groups heard the #-# tests. The order of presentation of the Man (M), Woman (W), and Child (C) tests was counterbalanced across the groups in the orders: MWC, WCM, CMW. Data for only the first two tests were analyzed (i.e., MW, WC, and CM, respectively). Thus, the total number of judgments by the Segregated test subjects was equivalent to that for the Mixed test subjects (90 judgments) and any effects of fatigue or familiarity were equally distributed across the three talkers for the Segregated tests.

Subjects. The data presented here for the /p-p/ conditions are those obtained in the previous study (Verbrugge et al., in press: Exp. II). Thirty-three subjects served in the Segregated /p-p/ tests (11 in each condition) and 19 subjects were tested on the Mixed /p-p/ test. For the tests on isolated vowels, 30 subjects were tested in the Segregated #-# test (10 per condition) and 16 subjects heard the Mixed #-# test. All subjects were paid volunteers from undergraduate psychology classes at the University of Minnesota. All were native speakers of English and most were natives of the upper midwest region.

Results

Errors in vowel identification were tabulated for each condition; an error was defined as the selection of a response other than that intended by the talker. The overall error rate for the four experimental conditions is shown in Figure 1. The main comparison of interest is between performance reported earlier for vowels in /p-p/ environment and performance on the isolated vowels. On the average, there were 17.0 percent errors on the Mixed /p-p/ test and 9.5 percent errors on the Segregated /p-p/ test. For the isolated vowels, on the other hand, there were 42.6 percent errors on the Mixed test and 31.2 percent

errors on the Segregated test. Errors summed over all nine vowels for each subject were submitted to a 2 × 2 analysis of variance for unequal cell frequencies. The main effects for talker variation (Mixed vs. Segregated) and consonantal context (/p-p/ vs. #-#) were both significant [$F(1,94) = 21.18$ and 125.17, respectively, $p < .01$]. However, no significant interaction between the two variables was found [$F(1,94) = 0.93$].

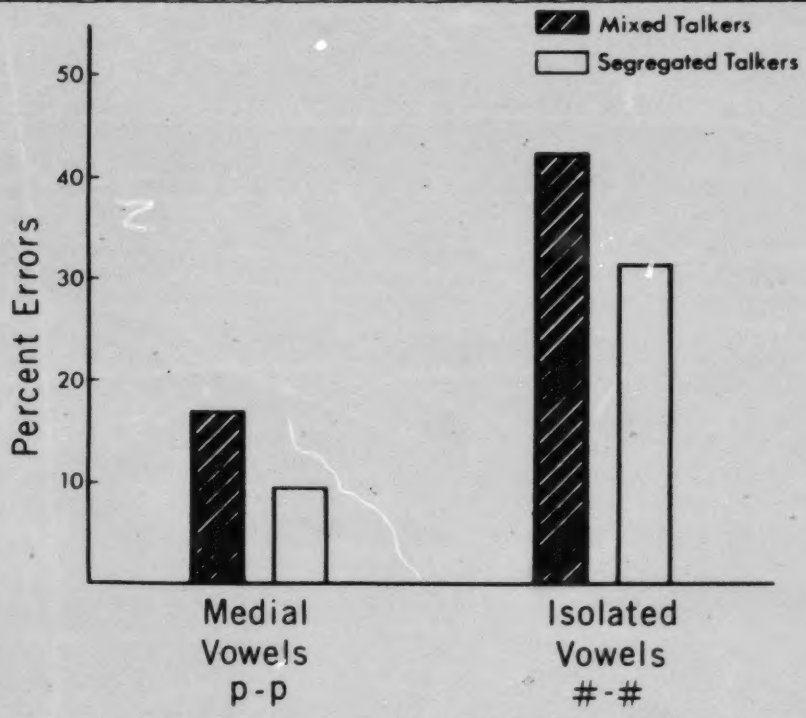


Figure 1: Overall percent errors for vowels in /p-p/ syllables and isolated vowels. Open bars show errors for Segregated Talker conditions; shaded bars show errors for Mixed Talker conditions.

These results indicate that while talker variation does contribute significantly to vowel identification errors for both medial vowels and isolated vowels, the presence or absence of consonantal context is by far the more important variable. Listeners misidentified approximately three times as many isolated vowel tokens as they did the corresponding medial vowels. Thus, it appears that the presence of a consonantal environment is much more critical for accurate vowel identification than is familiarity with the characteristics of the talkers' vocal tracts.

The hypothesis that consonantal environment contributes to perception of the vowel by providing cues for talker normalization was not supported. There was no interaction between the two major variables; the increased error rate due to the absence of consonantal context was almost as great when the talker was constant (an increase of 22 percent) as it was when talkers varied from token to token (an increase of 26 percent). We can conclude that the efficacy of the /p-p/ context in aiding vowel identification is directly involved with specification of vowel identity.

A vowel-by-vowel analysis of the identification errors for the four experimental conditions is presented in Table 1. (Confusion matrices for the /p-p/ and #-# tests are presented in Appendices A-1, A-2, A-3, and A-4.) It is readily apparent that for every vowel category, in both Mixed and Segregated Talker conditions, there were more errors for the isolated vowel than for the corresponding vowel in the /p-p/ frame. This is strong evidence that the lack of familiarity with a talker's vocal tract is far less detrimental to accurate perception of vowels than is the absence of information provided by a consonantal environment.

TABLE 1: Experiment I: Identification errors (in percent) for each intended vowel in four experimental conditions. Error rates excluding /ɑ/-/ɔ/ confusions are given in parentheses. (See Footnote 1.)

Intended Vowel	Segregated Talkers		Mixed Talkers	
	#-#	/p-p/	#-#	/p-p/
i	16	< 1	26	1
ɪ	14	4	23	2
ɛ	46	12	62	27
æ	26	2	48	19
ɑ	64 (19)	23 (4)	61 (32)	20 (10)
ɔ	29 (14)	18 (2)	30 (10)	27 (3)
ʌ	42	8	63	15
ʊ	29	18	49	39
u	14	< 1	23	3
Overall errors	31% (25)	9% (6)	43% (38)	17% (13)

The data reveal differenees in the identifiability of particular isolated vowels. The pattern of errors is quite similar to that found for medial vowels; the vowels /i/, /ɪ/, and /u/ are most accurately identified while the more central vowels yield relatively more errors in identification. It should be noted, however, that even the former show error rates from 14 to 26 percent when they are presented without consonantal context, compared to less than 4 percent errors obtained for these vowels in the /p-p/ context.[1]

A more detailed analysis was undertaken to evaluate the consistency of these results. The percent errors obtained for each of the 45 tokens on the Mixed #-# test was compared to the percent errors obtained for the comparable

[1]The extremely high error rate for the vowel /ɑ/ is, in part, due to the considerable confusion between /ɑ/ and /ɔ/ in the dialect of the talkers. In Table 1 the percentages shown in parentheses for these two vowels represent the error rates excluding /ɑ/-/ɔ/ confusions; that is, a response was counted correct if the subject identified an intended /ɑ/ either as /ɑ/ or as /ɔ/, and likewise for an intended /ɔ/. Adjusted overall error rates also presented in Table 1 show that subtracting /ɑ/-/ɔ/ confusions has little effect on the relative differences among the four conditions.

token on the Mixed /p-p/ test. Isolated vowel tokens were misidentified more often than medial vowels in 39 out of 45 cases, while two pairs produced an equal proportion of errors. In only four cases did the /p-p/ token produce more errors than the comparable isolated vowel. Thus, we can conclude that the difference in error rates found between performance on medial and isolated vowels is consistent across individual tokens of the vowels as well as across vowel categories.

The overall results of the Segregated tests show that isolated vowels were identified far less accurately than were medial vowels, even when talker variation was absent. Error rates for the man, woman, and child on the Segregated #-# tests were 33, 26, and 32 percent, respectively. Comparable error rates for the Segregated /p-p/ tests, reported in Verbrugge et al. (in press), were 9, 6, and 11 percent, respectively. The differences show a relatively constant advantage of consonantal environment for all three talkers, despite some variability in overall intelligibility of the talkers.

In summary, it is clear that consonantal environment contributes in a major way to the identification of vowels. We reach this conclusion whether we regard the data in terms of overall results, the results for particular vowel categories, for individual tokens, or for individual talkers. Isolated vowels are much more poorly identified than vowels embedded in the /p-p/ context.

Acoustical Analysis

The results of this experiment indicate that isolated, steady-state vowels are poor stimuli from the standpoint of the perceiver. The possibility remains, however, that the perceptual problem in identifying isolated vowels is a result of the way the talkers produced them. Phonetically untrained talkers may be unable to produce specified tokens of vowels reliably in isolation. Acoustical analysis of the vowel utterances by our panel of talkers was undertaken to investigate this possibility.

Center frequencies of the first three speech formants and the duration of the vocalic portion of each syllable were determined from spectrograms and spectral sections produced on a Voiceprint Sound Spectrograph. Recordings of tokens produced by women and children were reproduced at half-speed for spectrographic analysis; obtained frequency values were doubled to determine the actual formant frequencies of these tokens. Spectral sections were made at the point of nearest approach to the steady state. (If the vowel was diphthongized by the talker, measurements were obtained from the initial part of the vocalic portion of the syllables.) Two judges, working independently, determined the center frequency values for the speech formants to the nearest 25 Hz. Frequencies reported represent an average of the values obtained by the two judges. In addition, measurements of the duration of the first-formant periodic energy were made.[2]

---

[2] For many isolated vowels and some vowels in /p-p/ frames, the offset of periodic energy preceded offset of higher formant energy considerably. However, the rank order of vowels within each listening condition was the same even when the duration of higher formant energy was considered. Thus, the conclusions discussed in the text are valid for both measures of duration.

Measurements were obtained for the 45 tokens of the Mixed Talker /p-p/ test and the 45 isolated vowel tokens in the Mixed #-# test. In addition, measurements were obtained for the remaining six isolated vowels spoken by each talker that were not incorporated in the Mixed #-# test. Thus, one token of each of nine isolated vowels was measured for each of 15 talkers. For the Segregated tests, one token of each of the nine isolated vowels was selected randomly from each of the three talkers' tests. For comparison, the /p-p/ token that corresponded to each selected isolated vowel was also analyzed.

TABLE 2: Average frequency values (in Hz) for the first three speech formants of the nine isolated vowels, averaged over five talkers in each group.

		i	ɪ	ɛ	æ	ɑ	ɔ	ʌ	ʊ	u
F_1	M	355	447	635	737	757	672	685	497	337
	W	385	482	747	820	843	692	815	577	435
	C	357	580	755	885	1030	770	895	557	500
F_2	M	2245	1960	1790	1697	1220	942	1167	1092	1042
	W	2792	2325	2157	2110	1372	1312	1525	1399	1175
	C	3335	2710	2485	2685	1565	1350	1630	1340	1150
F_3	M	2937	2575	2510	2445	2347	2453	2307	2352	2165
	W	3482	3060	2960	2900	2915	2875	2847	2815	2735
	C	3880	3630	3765	3680	3700	3540	3725	3613	3150

Looking first at the analysis of the isolated vowels spoken by the full panel of talkers, we can ask whether the poor identification (43 percent errors) was due to the talkers' inability to produce isolated vowels reliably. Table 2 presents the average values of the first three speech formants for the men, women, and children. In Figure 2 the average values for the first and second speech formants are plotted in a two-dimensional "vowel space." On the average, our talkers' productions of the vowels in isolation were systematic in distribution and corresponded closely in formant values to vowels sampled by other investigators (Peterson and Barney, 1952; Tiffany, 1959; Stevens and House, 1963). The formant frequencies showed systematic elevations from men to women to children, reflecting a general decrease in the size of these talkers' vocal tracts.

Individual tokens of isolated vowels corresponded closely to values reported in previous studies except for tokens of the vowel /ɔ/ by all talkers, tokens of /ɛ/ spoken by the men and women, three tokens of /æ/ spoken by children, and one token of /ʊ/ spoken by a woman. The deviation in /ɔ/ tokens represents a dialectal difference between our talkers and those recorded by Peterson and Barney (1952). Stevens and House (1963) did not report data for this vowel.

The next question of interest is whether the panel's productions of isolated vowels differed greatly from their corresponding productions of vowels in the /p-p/ consonantal frame. To answer this question, we compared the tokens actually used in the two Mixed Talker tests. Figure 3 presents the average values of F_1 and F_2 for the medial vowels and isolated vowels, pooled across

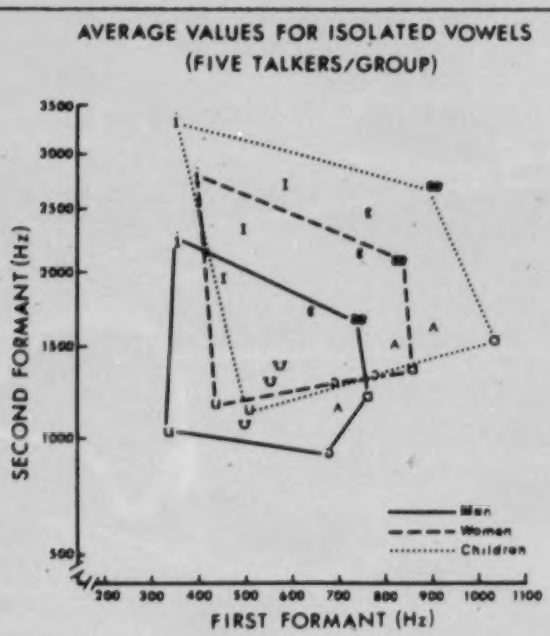


Figure 2: Average Formant 1/Formant 2 values for isolated vowels spoken by men, women, and children (five talkers in each group).

men, women, and children. The vowels on the two tests occupied almost the same area in F_1/F_2 space. The second formant of the medial vowels showed a slight migration toward the center of the space. This is an expected result of coarticulation (where formants fail to reach a steady-state target) and is in accord with results reported by Stevens and House (1963) for vowels produced between consonants with labial and labio-dental place of articulation. A. Tiffany (1959) noted, this reduces the acoustic contrast among vowels spoken in a consonantal frame in comparison to isolated vowels. However, the perceptual data demonstrate that identifiability cannot be predicted from the spread of steady-state formant measurements; medial vowels were perceptually much more distinct than vowels in isolation (83 vs. 57 percent correct identifications).

The two sets of vowels were very similar in formant frequencies, in both the central tendency and the variability of values for each vowel. Even so, there were a few individual tokens that deviated markedly from the central tendencies. It is of interest whether the considerably greater error rate for isolated vowels over that obtained for medial vowels can be attributed primarily to the misidentification of tokens that were produced in a deviant manner.

One way to answer this question is to look at those pairs of tokens that contributed most to the difference obtained in the perceptual tests. For nine comparison pairs, errors for the isolated vowel exceeded those for the medial vowel by more than 50 percent of the opportunities for error. It might be supposed that the formant frequencies of these isolated vowel tokens would show the greatest deviation from the average values and from values for the comparable

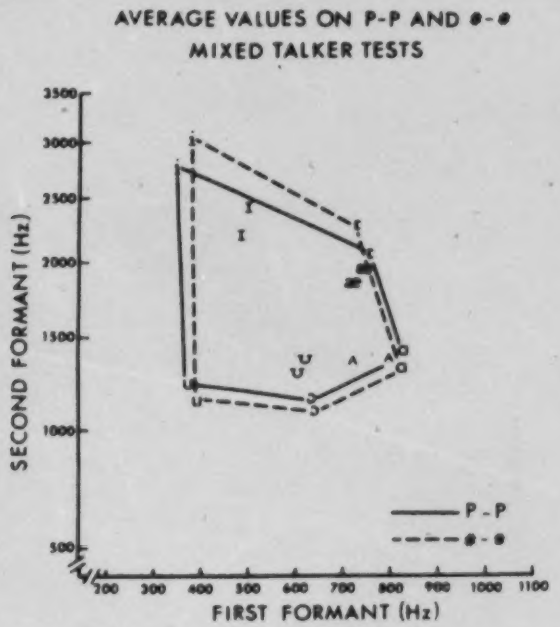


Figure 3: Average Formant 1/Formant 2 values for vowels in /p-p/ syllables (solid lines) and vowels in isolation (dashed lines). Values were computed over the five tokens of each vowel in each Mixed Talker test.

medial vowel. This is not the case, however, as may be seen from Figure 4, which shows the nine vowel pairs. For some of these pairs, the first- and second-formant values for both the isolated and medial vowels fell within the range of variation for the appropriate vowel category. For the vowels /æ/, /ɑ/, and /ɪ/ both isolated and medial vowels were displaced from their typical positions. Finally, for the vowels /ʊ/, /ɛ/, and one pair of /ʌ/, the isolated vowel might be considered <u>less</u> confusable acoustically than its counterpart in medial position. Thus, there seems to be no close correspondence between perceptual confusability and acoustic deviation from some expected (target) value.

This does not mean, of course, that variations in formant frequency positions have no effect on perception. There were a few pairs of tokens that were "misarticulated" on both the /p-p/ and #-# tests and that contributed relatively greater numbers of errors in identification. (For example, one woman's production of /ʊ/ was quite deviant on the medial vowel test, as well as on the isolated vowel test. Listeners made 38 and 100 percent errors on the isolated and medial tokens, respectively.) However, with respect to the present comparison, the salient point is that deviation in formant structure cannot account for the large and consistent differences between perceptual tests of isolated vowels and vowels in a fixed consonantal frame.

Measurements of formant frequencies of tokens from the Segregated Talker tests corroborate the results for the Mixed Talker tests. Since measurements were made for only a sample of the total set of items, we cannot be sure that deviations in the production of isolated vowels were not responsible for their inferiority as perceptual targets. However, the tokens that were measured gave

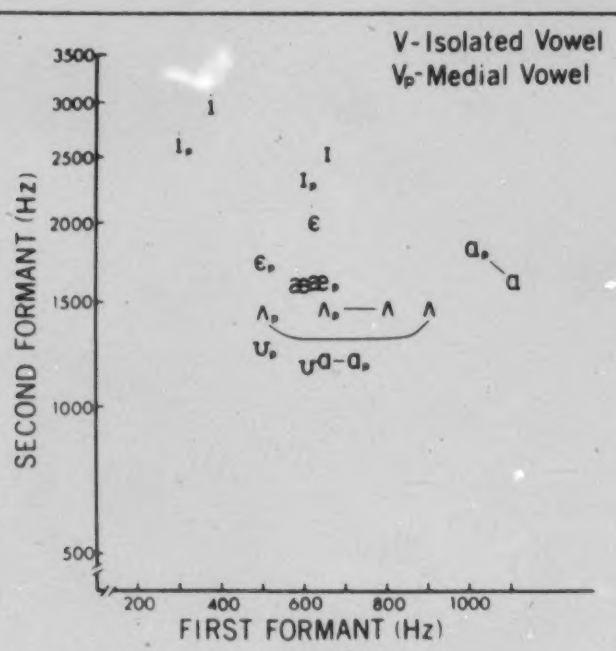


Figure 4: Formant 1/Formant 2 values for the nine pairs of vowels on the Mixed Talker tests that contributed most to the difference in identification errors. Vowels in /p-p/ syllables are indicated by the subscript p.

no indication that the three talkers produced the isolated vowels less consistently than they did the medial vowels. A comparison of pairs of tokens showed that isolated and medial vowels were similar in all but a few cases. Deviations from the normal range of formant values were as likely to be obtained for a randomly selected medial vowel as they were for a randomly selected isolated vowel. Thus, the consistent advantage found in perceptual tests for medial vowels over isolated vowels, for all three talkers and all nine vowel categories, cannot be attributed to deviant formant frequencies of isolated vowels.

While there was no indication of large differences in the formant structure of the vowels in isolation and those in syllables spoken in citation form, these two sets of tokens did differ considerably in terms of overall duration. Table 3 gives the average duration of the voiced first formants of isolated and medial vowels in Segregated and Mixed Talker tests. The isolated vowels were much longer on the average than were the medial vowels. However, a more important consideration is the relative durations of the vowels in the two sets. More specifically, are the relative durations of isolated vowels different from those typically found for vowels in consonantal context?

The relative durations of vowels in /p-p/ frames were similar to the values reported by Peterson and Lehiste (1960) and House and Fairbanks (1953). The vowels, /ɪ/, /ɛ/, /ʌ/, and /ʊ/ were the shortest in duration; /i/ and /u/ were intermediate; and /ɑ/, /ɔ/, and /æ/ were the longest vowels. The only exception to this in our data was the vowel /u/ in the Segregated /p-p/ test, for which the average duration was considerably shorter than that reported by other researchers.

TABLE 3: Experiment I: Average durations (in msec) of the vocalic portion of tokens in four experimental conditions. Asterisks indicate deviant lengths (see text).

Intended Vowel	Segregated Talkers[a]		Mixed Talkers[b]	
	#-#	/p-p/	#-#	/p-p/
i	315	128	326*	148
ɪ	228	108	198	138
ɛ	226	111	245*	136
æ	328	194	256	204
ɑ	313*	179	237	177
ɔ	303*	186	251	186
ʌ	246	116	184	138
ʊ	242	124	259*	131
u	311	109*	237	159
Overall errors	279.1	139.4	243.7	157.4

[a] Averages based on three randomly selected tokens of each vowel, one from each of the three talkers.

[b] Averages based on five tokens, each spoken by a different talker.

As Table 3 indicates, relative durations for the isolated vowels were similar to those for medial vowels with the following exceptions: for the Mixed #-# test, the vowels /i/, /ɛ/, and /ʊ/ showed longer relative durations than they did in consonantal context. For the Segregated test, the vowels /ɑ/ and /ɔ/ showed shorter relative durations than their counterparts in consonantal frames.

The atypical durations of these isolated vowels cannot account for the consistent advantage of medial vowels over isolated vowels for every vowel category in the perceptual tests. Even for the deviant vowels, the confusion patterns showed no consistent trend toward responses that would be predicted on the basis of the deviant durations. (See Appendices A-3 and A-4 for confusion matrices.)

Discussion

In this study we found that vowels produced in a fixed consonantal environment were identified with much greater accuracy than were comparable steady-state vowels produced in isolation. This was true both when variation due to talker differences was present and when it was not. Thus, the experiment provides no evidence that coarticulated consonants facilitate identification by enabling the listener to recalibrate for each new talker. Coarticulated consonants are integral to the specification of vowels whether a talker is familiar or not.[3]

[3] It has been suggested that the relatively poor performance on the isolated vowels might be due to the lack of correspondence between the stimuli and the

Acoustical analyses were undertaken to investigate the possibility that untrained talkers fail to adopt consistent targets for vowels in isolation, resulting in a highly unreliable signal for perception. Although there were systematic acoustic differences between vowels produced in consonantal environment and those produced in isolation, the large and consistent increases in confusability among isolated vowels over those obtained for medial vowels could not be explained by increases in the acoustic similarity of vowel categories when defined by formant frequencies. Nor could these differences be attributed to differences in the relative durations of the vowels in isolation and in context. It is interesting to note that medial vowels tend to be more similar to each other than comparable isolated vowels in terms of the cross-sectional acoustic parameters that have traditionally been used to differentiate vowel classes. This is additional support for the view that static descriptions of vowels are inadequate for capturing perceptually relevant aspects of the acoustic signals. Our results lead us to conclude that the acoustic information for vowel identity, like that for consonants, is specified in the dynamic configuration of the syllabic pattern as a whole.

In this study, the consonantal environment in which the vowels were produced was constant across all tokens. Thus, the listeners knew beforehand the identity of two of the three phonemes in each test token. It is possible that this knowledge (rather than the presence of formant contours) was the source of superior identification for medial vowels. It would be of limited interest if consonantal environment aided in vowel identification only in this circumstance, since it is not generally the case that listeners have advance knowledge of consonantal identity in natural listening conditions. We therefore undertook an additional experiment to test the effects of a varying consonantal environment on the identification of medial vowels.

EXPERIMENT II: PERCEPTION OF VOWELS IN CVC SYLLABLES

We wanted to determine whether a consonantal context that varies from trial to trial (and is therefore unpredictable by the listener) provides

---

orthographic representation of the alternatives provided on the response forms. For both /p-p/ and #-# conditions, subjects were required to respond by selecting the appropriate /p-p/ syllable, for example, peep and pip. Thus, subjects in the #-# condition had to "decode" the orthography to match the isolated vowel, whereas subjects who heard medial vowels had only to match the orthographic syllable to the perceived syllable. Since the preparation of this manuscript, we have used different response forms for both /p-p/ and #-# tests. The symbols on the response forms corresponded to vowels in isolation, for example, EE, IH, and EH, and subjects were given practice to make sure they could use the symbols appropriately. Results of these studies, when compared to those from conditions using the syllable response alternatives, showed no difference in performance for the isolated vowels. On the other hand, errors for vowels in /p-p/ syllables were somewhat greater when we used the isolated vowel symbols. However, identification of medial vowels was still significantly better than for isolated vowels. Further studies of the effects of different response forms are underway and will be reported in a subsequent article. We feel quite confident that the large and consistent differences found in the present study were due primarily to perceptual effects.

important information for vowel identification. We again included conditions where the talkers varied from trial to trial (Mixed) and where the same talker produced all tokens on a particular test (Segregated), in order to investigate the possible interaction between talker variation and knowledge of consonantal context.

Method

Stimulus materials. The C-C test syllables were composed from six stop consonants, /p, t, k, b, d, g/, and the nine vowels used in Experiment I. A panel of four adult males, four adult females, and four children (a subset of the 15 talkers used in Experiment I) each produced six tokens for the Mixed Talker condition, resulting in a test series of 72 syllables. Within this series, each vowel occurred 8 times and each initial and final consonant occurred 12 times. Consonants and vowels were paired such that each vowel was preceded and followed by each consonant at least once. (Both symmetrical and nonsymmetrical pairings were used; for example, syllables such as /t-t/ and /d-t/ both appeared in the test series.) The assignment of syllables to talkers was random with the constraint that a talker did not produce the same vowel more than once, nor the same initial consonant more than twice.

The talkers read the test syllables from cards on which they were printed in standard English orthography, except in cases where no unambiguous English spelling existed. For these items, key words were provided beneath the test syllables to indicate that pronunciation of the vowel. All test stimuli were recorded using the equipment and procedures described in Experiment I.

The 72 test syllables were arranged in an order of presentation with the following restrictions: (1) the same intended vowel did not occur more than twice consecutively, (2) there was an equal number of tokens of each intended vowel in the first and second half of the test, (3) the same initial consonant did not occur more than twice consecutively, (4) tokens produced by the same talker were separated by not less than six tokens, and (5) each talker occurred equally often in the first and second half of the series. For the Segregated Talker tests, the same three talkers were recorded as in the Segregated tests in Experiment I. Each talker recorded the entire list of 72 syllables in the same order as for the Mixed Talker test.

Procedure. Listening tests were administered to small groups of subjects using the equipment and procedures described in Experiment I. Listeners responded on score sheets printed with columns of key letters representing each of the nine vowels. Above each column, key words containing these letters were printed as follows: "s<u>i</u>n s<u>u</u>m s<u>a</u>nd s<u>ee</u>n sh<u>o</u>p s<u>e</u>nt s<u>oo</u>n s<u>aw</u> sh<u>ou</u>ld." The key letters in the columns were preceded and followed by blank lines. Before the listening test, the experimenter pronounced each key word followed by its vowel in isolation. Special attention was drawn to the key letters that represented the vowel /ʊ/.

Subjects in the Mixed Talker condition were required to identify only the vowel in each syllable. They did this by circling, for each syllable, the key letter(s) that symbolized the perceived vowel. Listeners heard the entire test series twice for a total of 144 judgments per subject.

Three groups of subjects were tested in the Segregated Talker condition. All three groups were required to identify only the vowel in the syllables, and they did so in the same way as the subjects in the Mixed Talker condition. As in Experiment I, each group of subjects heard the three talkers in one of three orders: MWC, WCM, or CMW. Again, data for only the first two tests were analyzed, making the number of judgments per subject equal to that for the Mixed Talker tests (i.e., 144 judgments per subject). Subjects in all conditions were told that some of the test syllables were real words and that some were nonsense syllables, but that they were to ignore meaning and respond only on the basis of the sound of the syllables.

Subjects. All subjects were paid volunteers obtained from undergraduate psychology courses at the University of Minnesota. All were native speakers of English and most were natives of the upper midwest region. Twenty-two subjects served in the Mixed Talker condition. Twenty-four subjects were tested in the Segregated Talker condition, eight with each of the three counterbalanced orders.

Results and Disucssion

Table 4 presents the overall error rates for the two conditions of this experiment along with the results of Experiment I for comparison. There was no significant difference between the error rates for the Segregated Talker condition (22.9 percent) and the Mixed Talker condition (21.7 percent) [$t(44\ df) = 0.43$].

TABLE 4: Overall identification errors (in percent) for Experiments I and II.

		Segregated Talkers	Mixed Talkers
Experiment I	/p-p/ Test	9.5	17.0
	#-# Test	31.2	42.6
Experiment II	C-C Test	22.9	21.7

The major question of interest was whether consonantal context aids vowel identification even when the context is unpredictable. The results for the C-C test syllables may be compared with those found in Experiment I for /p-p/ syllables and isolated vowels (cf. Table 4). For the Mixed Talker condition, vowels in C-C syllables were identified with significantly greater accuracy than were comparable isolated vowels, as tested by a median test: χ^2 (1 df) = 18.24, $p < .01$. The overall error rate of 21.7 percent for C-C syllables was not significantly greater than the 17 percent errors found for vowels in /p-p/ syllables [χ^2 (1 df) = .23]. Thus, the results for the Mixed Talker condition are clear; both fixed and variable consonantal frames produced a dramatic improvement in vowel identifiability in contrast to isolated vowels. The advantage of a consonantal environment obtains even when the identity of the consonants is not known in advance by the listeners.[4]

[4] It is worth noting that tokens by the subset of 12 talkers used in the C-C test yielded 20 percent errors on the /p-p/ test. Thus, if anything, errors

The overall results for the Segregated Talker condition were less conclusive. Vowels in C-C syllables were, on the average, better identified than isolated vowels: χ^2 (1 df) = 6.08, $p < .02$. However, unlike the Mixed Talker results, listeners did not identify vowels in C-C syllables as accurately as vowels in /p-p/ syllables [χ^2 (1 df) = 25.6, $p < .01$]. The error rate for the Segregated C-C test appears to be idiosyncratic in that there was no advantage over the comparable Mixed Talker condition. (For the /p-p/ and #-# tests, the advantage of Segregated test over Mixed test was 8 and 12 percent, respectively.)

Table 5 presents the errors for each vowel category in the two C-C conditions. (Confusion matrices are given in Appendices A-5 and A-6.) Results for individual vowel categories in the Mixed Talker condition (right-hand column) verified the pattern found for overall errors. In comparison with the data for the Mixed #-# test (Table 1), vowels of each category, with the exception of /ɔ/, were identified with greater accuracy when they were spoken in a variable consonantal frame than when they were spoken in isolation.

TABLE 5: Experiment II: Identification errors (in percent) for each intended vowel in two experimental conditions.

Intended Vowel	Segregated Talkers	Mixed Talkers
i	8	6
ɪ	12	17
ɛ	14	24
æ	13	15
ɑ	41 (15)	31 (7)
ɔ	44 (10)	37 (11)
ʌ	11	18
ʊ	46	39
u	17	8
Overall errors	23% (17)	22% (16)

Results for individual vowel categories in the Segregated Talker tests (left-hand column) showed an unexpectedly high error rate for back vowels, /ɑ/, /ɔ/, /ʊ/, and /u/, for all three talkers. Errors on these vowels account for the lack of an overall advantage in the Segregated condition over the Mixed condition with C-C syllables. We currently have no explanation for this result.

The results of this experiment support the claim that consonantal context aids in the specification of vowel identity by providing important acoustic information to the listener. Even when the consonants are not known in advance, listeners are much more accurate in identifying medial vowels in CVC syllables than they are in identifying isolated steady-state vowels.[5] The acoustic

in the C-C study are probably overestimated relative to the results one might expect for a test including all 15 talkers.

[5] In a separate study, similar results were found when subjects were asked to identify both the consonants and the vowel in each test syllable. Errors in

effects of coarticulation carry substantial information about a medial vowel, which aids in vowel identification whether or not the listener has prior knowledge of the consonants' identity.[6]

SUMMARY AND CONCLUSIONS

In Experiment I, perceptual tests of vowels produced in isolation and in a fixed CVC context by the same talkers demonstrated that providing a consonantal environment increases the likelihood of correct identification of the intended vowel. This was true both when talker variation was present and when it was not; the advantage of consonantal context was independent of talker variation. Of the two factors investigated, consonantal context was much more important than talker variation in determining listeners' identification of vowels. The increment in error for isolated vowels in comparison to the medial vowels was more than three times greater than the increment attributable to unpredictability of talker.

We considered what might account for the difference in intelligibility between vowels in /p-p/ environment and in isolation. We concluded that the poor

---

vowel identification averaged 29 percent. Thus, even with the additional task of identifying the consonants, error rates were substantially lower than when listeners were required to identify vowels in isolation.

[6] Two aspects of the design of the C-C tests make further interpretation of the results problematic. First, although each consonant appeared equally often, the occurrences of consonants in initial and final position were not balanced across vowels, nor were equal numbers of consonants contributed by different talkers in the Mixed test. As a result, we cannot make precise statements about the relative advantages of fixed and variable contexts, about the interaction of context with talker variation, or about the relative effects of different consonants on the identifiability of coarticulated vowels. A second problem concerns a possible interaction between vowel categories and prior familiarity with particular test items. Many of the C-C syllables are words that are familiar to the listeners. If this factor has a major effect on the perception of vowels in tasks like ours (in spite of the closed response set and the instructions to ignore meaning), the superior recognition of C-C syllables might have little to do with the type of acoustic information made available. If so, one might expect that listeners would do far better on syllables that formed words than on those that were nonsense syllables. Of the 72 C-C syllables included in the present experiment, 38 were English words. The overall error rate for these tokens in the Mixed Talker test was 15 percent, compared to a 25 percent error rate for the 34 remaining C-C syllables. While this suggests that linguistic experience is a factor in vowel identification under these conditions, two further observations should be made. First, both error rates are well below that obtained for isolated vowels. Thus, if experience is a factor at all, it is probably secondary to the presence of phonetic context. Second, the error rates for the real words and nonsense syllables are difficult to interpret, since the fraction of C-C syllables that are real words varies with different vowel categories. The analysis is further complicated by intrinsic differences in perceptual difficulty among the nine vowels and by differences among the C-C syllables in orthographic representation.

intelligibility of isolated vowels could not be attributed to the talkers' failure to produce these vowels in a consistent manner or to their adoption of aberrant formant frequencies. Measurements showed that formant frequency values and relative durations of isolated vowels were generally quite similar to those of vowels in the consonantal frame. The relative intelligibility of a token cannot be estimated very precisely from its position in the space defined by the two formants, a fact also noted by Peterson and Barney (1952).

The second experiment showed that consonantal context aids vowel identification even when the consonant frame varies unpredictably. Vowels produced in randomly varying stop-consonant environments were identified more accurately than were isolated vowels both when the talker was fixed within a test block and when talkers, as well as context, varied unpredictably.

These results are surely puzzling if one makes the assumption that target frequencies of the formants alone could fully specify the vowels. If that were so, an isolated quasi-steady-state utterance ought to be an optimal signal for perception. It is true that synthetic steady-state vowels based on these formant parameters are fairly intelligible to naive listeners and may be identified quite consistently by experienced listeners (Delattre, 1951). Moreover, in the domain of automatic speech recognition, some success has been achieved with a static model of the vowel. Gerstman (1968) devised an algorithm based on frequencies of the first and second formants of /h-d/ syllables recorded from 76 talkers by Peterson and Barney (1952). Gerstman's algorithm sorted nine vowels in this set with only 2.5 percent error, less than was made by human listeners. From such a result, one might infer that target formant frequencies can unambiguously specify the vowels of English as produced by a variety of talkers.

However, as we have seen, this conception of the vowel cannot be reconciled easily with certain facts of perception. Vowels in isolation were poor signals from the perceiver's standpoint, even though talkers adopted targets that differed little from those attained in citation-form /p-p/ syllables. Thus, we may suspect that no single cross section through the syllable can fully specify the vowel. This inference is consistent with previous studies in the phonetic literature, to which we have referred. It is also relevant, in this context, to mention the results of an experiment by Bond (1975) on perception of vowels created by iteration of a single cycle from steady-state vowel tokens. Perception of such vowels by naive listeners was even less reliable than the results we obtained for unedited isolated vowels. If target frequencies alone were fully adequate to specify the vowels, it is difficult to understand these results.[7]

We are led to conclude that cues that are ordinarily regarded as consonantal contribute regularly to the perception of the vowel. We suspect that much vowel information is contained in formant transitions, as Lindblom and Studdert-Kennedy (1967) suggested some time ago. Whatever the nature of the contribution consonantal environment makes to the identification of a vowel, the data we have reviewed point to the general conclusion that no single temporal cross section of a syllable conveys as much vowel information to a perceiver as is given in the dynamic contour of the formants. From the standpoint of perception, it

[7] The implications of the specification of vowels in terms of idealized "targets" is explored further in Shankweiler, Strange, and Verbrugge (in press).

would seem that the definition of a vowel ought to include a specification of how the relevant acoustic parameters change over time. While listeners may be trained to identify steady-state tokens accurately (Lehiste and Meltzer, 1973), there is no reason to believe that the processes involved in this activity are the same as those typically used for understanding speech in natural situations.

Finally, these results may have implications for understanding the vocal-tract normalization problem. Attempts to specify vowels across talkers have usually taken as their basic data, the formant frequency values of a single cross section of a syllable. Our research indicates that the human perceptual system is ill-equipped to deal with such data. It would seem fruitful to renew the search for invariants across talkers utilizing information defined over the time course of at least a syllable.

REFERENCES

Abramson, A. S. and F. S. Cooper. (1959) Perception of American English vowels in terms of a reference system. Haskins Laboratories Quarterly Progress Report QPR-32, Appendix 1.

Bond, Z. S. (1975) Identification of vowels excerpted from context. J. Acoust. Soc. Am., Suppl. 57, S24(A).

Delattre, P. C. (1951) The physiological interpretation of sound spectrograms. Publications of the Modern Language Association of America 66, 864-875.

Fairbanks, G. and P. Grubb. (1961) A psychophysical investigation of vowel formants. J. Speech Hearing Res. 4, 203-219.

Fourcin, A. J. (1968) Speech source inference. IEEE Trans. Audio Electroacoust. AU-16, 65-67.

Fujimura, O. and K. Ochiai. (1963) Vowel identification and phonetic contexts. J. Acoust. Soc. Am. 35, 1889(A).

Gerstman, L. H. (1968) Classification of self-normalized vowels. IEEE Trans. Audio Electroacoust. AU-16, 78-80.

Harris, C. M. (1953) A study of the building blocks in speech. J. Acoust. Soc. Am. 25, 962-969.

House, A. S. and G. Fairbanks. (1953) The influence of consonant environment upon the secondary acoustical characteristics of vowels. J. Acoust. Soc. Am. 25, 105-113.

Jones, D. (1956) An Outline of English Phonetics. (Cambridge, England: W. Heffer).

Lehiste, I. and D. Meltzer. (1973) Vowel and speaker identification in natural and synthetic speech. Lang. Speech 16, 356-364.

Liberman, A. M. (1970) The grammars of speech and language. Cog. Psychol. 1, 301-323.

Liberman, A. M., F. S. Cooper, D. P. Shankweiler, and M. Studdert-Kennedy. (1967) Perception of the speech code. Psychol. Rev. 74, 431-461.

Lindblom, B. E. F. (1963) Spectrographic study of vowel reduction. J. Acoust. Soc. Am. 35, 1773-1781.

Lindblom, B. E. F. and M. Studdert-Kennedy. (1967) On the role of formant transitions in vowel recognition. J. Acoust. Soc. Am. 42, 830-843.

Millar, J. B. and W. A. Ainsworth. (1972) Identification of synthetic isolated vowels and vowels in h-d context. Acustica 27, 278-282.

Öhman, S. E. G. (1966) Coarticulation of VCV utterances: Spectrographic measurements. J. Acoust. Soc. Am. 39, 151-168.

Peterson, G. E. and H. L. Barney. (1952) Control methods used in a study of the vowels. J. Acoust. Soc. Am. 24, 175-184.

Peterson, G. E. and I. Lehiste. (1960) Duration of syllable nuclei in English. J. Acoust. Soc. Am. 32, 693-703.

Potter, R. K. and J. C. Steinberg. (1950) Toward the specification of speech. J. Acoust. Soc. Am. 22, 807-823.

Rand, T. C. (1971) Vocal tract size normalization in the perception of stop consonants. Haskins Laboratories Status Report on Speech Research SR-25/26, 141-146.

Schatz, C. (1954) The role of context in the perception of stops. Language 30, 47-56.

Shearme, J. N. and J. N. Holmes. (1962) An experimental study of the classification of sounds in continuous speech according to their distribution in the formant 1 - formant 2 plane. In Proceedings of the Fourth International Congress of Phonetic Sciences, ed. by A. Sovijärvi and P. Aalto. (The Hague: Mouton), pp. 234-240.

Shankweiler, D. P., W. Strange, and R. R. Verbrugge. (in press) Speech and the problem of perceptual constancy. In Perceiving, Acting, and Knowing: Toward an Ecological Psychology, ed. by R. Shaw and J. Bransford. (Hillsdale, N. J.: Lawrence Erlbaum Assoc.).

Stevens, K. N. and A. S. House. (1963) Perturbations of vowel articulations by consonantal context: An acoustical study. J. Speech Hearing Res. 6, 111-128.

Strange, W., R. R. Verbrugge, and D. Shankweiler. (1974) Consonant environment specifies vowel identity. Haskins Laboratories Status Report on Speech Research SR-37/38, 209-216.

Tiffany, W. R. (1959) Nonrandom sources of variation in vowel quality. J. Speech Hearing Res. 2, 305-317.

Verbrugge, R. R., W. Strange, D. P. Shankweiler, and T. R. Edman. (in press) What information enables a listener to map a talker's vowel space? J. Acoust. Soc. Am. [Also in Haskins Laboratories Status Report on Speech Research SR-45/46 (this issue).]

APPENDIX A: CONFUSION MATRICES

Tables report the frequency with which each intended vowel x was identified as response alternative y. In addition, summary statistics for each condition are provided: the percent error for each intended vowel, the overall percent error, and the number of listeners (N).

TABLE A-1: Vowels in /p-p/ syllables: Mixed Talker condition.[a]

Intended vowel	Response										Percent error
	i	ɪ	ɛ	æ	ɑ	ɔ	ʌ	ʊ	u	None	
i	188		1						1	1	1.1
ɪ		187	1			2					1.6
ɛ			139	47	3			1			26.8
æ			33	154		2				1	18.9
ɑ					152	19	17	2			20.0
ɔ				1	46	138	1	4			27.4
ʌ					18	5	161	6			15.3
ʊ		8			2		47	116	16	1	38.9
u							2	3	185		2.6

[a]Overall percent error = 17.0 percent; N = 19.

TABLE A-2: Vowel in /p-p/ syllables: Segregated Talker condition.[a]

Intended vowel	Response										Percent error
	i	ɪ	ɛ	æ	ɑ	ɔ	ʌ	ʊ	u	None	
i	329	1									0.3
ɪ	3	318	4				2	2		1	3.6
ɛ	1		290	20	4	7	5			3	12.1
æ			5	324		1					1.8
ɑ				7	255	62	4	2			22.7
ɔ					55	269	2	4			18.5
ʌ					11	9	305	4		1	7.6
ʊ						29	19	272	10		17.6
u							1	2	327		0.9

[a]Overall percent error = 9.5 percent; N = 33.

TABLE A-3: Isolated vowels: Mixed Talker condition.[a]

Intended vowel	Response										Percent error
	i	ɪ	ɛ	æ	ɑ	ɔ	ʌ	ʊ	u	None	
i	119	30	6					1	4		25.6
ɪ	2	124	19			3	6	1	1	4	22.5
ɛ	1	2	61	64	2	6	10	5	3	6	61.9
æ		2	51	84	3	10	1	6	2	1	47.5
ɑ	1		1	20	62	47	21	2		6	61.3
ɔ		1	2	2	18	112	17	6	1	1	30.0
ʌ		1		6	32	31	60	22	4	4	62.5
ʊ		1	5	3	1	15	48	81	1	5	49.4
u	2		1	1		7	6	16	124	3	22.5

[a]Overall percent error = 42.6 percent; N = 16.

TABLE A-4: Isolated vowels: Segregated Talker condition.[a]

Intended vowel	Response										Percent error
	i	ɪ	ɛ	æ	ɑ	ɔ	ʌ	ʊ	u	None	
i	251	3	1	1		1	1	6	33	3	16.3
ɪ	5	259	21		1	3	3	1	4	3	13.7
ɛ	4	7	161	92	9	6	9	7		5	46.3
æ			48	221	3	18	3	2	3	2	26.3
ɑ			2	37	107	135	17	1	1		64.3
ɔ		1	1	12	43	214	19	6		4	28.7
ʌ		1	6	30	47	31	174	9		2	42.0
ʊ			3	4	3	10	51	214	12	3	28.7
u	8	1	1	3	1	2	3	22	258	1	14.0

[a]Overall percent error = 31.2 percent; N = 30.

TABLE A-5: Vowels in C-C syllables: Mixed Talker condition.[a]

Intended vowel	Response										Percent error
	i	ɪ	ɛ	æ	ɑ	ɔ	ʌ	ʊ	u	None	
i	331	7	5	1			1		5	2	6.0
ɪ	2	292	53	1			2	2			17.1
ɛ	3	20	269	31	2		21	3		3	23.6
æ			47	298		7					15.3
ɑ		4	2	6	242	85	6	4	1	2	31.3
ɔ	2	3	1	2	91	222	18	6	4	3	36.9
ʌ			21	5	14	4	289	17	1	1	17.9
ʊ	1	6	1		8	10	70	214	41	1	39.2
u	5				2		6	16	323		8.2

[a]Overall percent error = 21.7 percent; N = 22.

TABLE A-6: Vowels in C-C syllables: Segregated Talker condition.[a]

Intended vowel	Response										Percent error
	i	ɪ	ɛ	æ	ɑ	ɔ	ʌ	ʊ	u	None	
i	354	2	17		1			1	5	4	7.8
ɪ	4	339	35					1	1	4	11.7
ɛ	10	21	329	13	1		1		1	8	14.3
æ	2	1	28	333	2	7		1		10	13.3
ɑ		1	1	23	225	100	15	4	6	9	41.4
ɔ		1		11	130	217	4	10	4	7	43.5
ʌ			3	3	16	8	342	8		4	10.9
ʊ	2	4	2		10	1	53	209	91	12	45.6
u	1	1	1		5	2	5	48	318	3	17.2

[a]Overall percent error = 22.9 percent; N = 24.

What Information Enables a Listener to Map a Talker's Vowel Space?*

Robert R. Verbrugge,[+] Winifred Strange,[++] Donald P. Shankweiler,[+++] and Thomas R. Edman[++]

ABSTRACT

Prior experience with a talker's speech contributes little to success in vowel identification. Adult listeners avaraged only 12.9 percent errors on 15 vowels in /h-d/ syllables spoken in mixed order by 30 talkers (men, women, and children), and 17.0 percent errors on 9 vowels spoken in /p-p/ syllables by 15 talkers. When the /p-p/ test series was spoken by single talkers, errors decreased by less than half to 9.5 percent. Experience with known subsets of a talker's vowels did not significantly reduce errors on subsequent test tokens: following the point vowels (/i/, /ɑ/, /u/), errors averaged 12.2 percent on vowels in /h-d/ context and 15.2 percent in /p-p/ context; following three central vowels (/ɪ/, /æ/, /ʌ/), errors averaged 14.9 percent in /p-p/ context. Precursors mainly influenced listeners' response biases, rather than facilitating true improvements in vowel identifiability. These results did not support the hypothesis that point vowels provide listeners with unique information for normalizing a talker's "vowel space." Errors on vowels in rapid, destressed /p-p/

---

*A partial summary of these results was presented at the 87th meeting of the Acoustical Society of America, New York, 25 April 1974 (see Verbrugge, Strange, and Shankweiler, 1974; see also Shankweiler, Strange, and Verbrugge, in press). This article is to be published in the Journal of the Acoustical Society of America (1976).

[+]University of Minnesota, Minneapolis; currently at University of Michigan, Ann Arbor.

[++]University of Minnesota, Minneapolis.

[+++]Also University of Connecticut, Storrs.

Acknowledgment: This paper reports research begun during the academic year 1972-73 while D. Shankweiler was a guest investigator at the Center for Research in Human Learning, University of Minnesota, Minneapolis. The work was supported by grants to the Center and to Haskins Laboratories from the National Institute of Child Health and Human Development, by grants awarded to D. Shankweiler and J. J. Jenkins by the National Institute of Mental Health, and by a fellowship to R. Verbrugge from the University of Michigan Society of Fellows. We wish to thank Kevin Jones, Kathleen Briggs, Robert Jenkins, and Mark Jaffe for their assistance in the experimental work, Keith Smith for his helpful advice on data analysis, and James Jenkins for his advice and encouragement throughout this research.

[HASKINS LABORATORIES: Status Report on Speech Research SR-45/46 (1976)]

syllables (excised from sentence context) averaged 23.8 percent. Errors jumped to 28.6 percent when point-vowel precursors were introduced, while presentation of syllables in the original sentences reduced errors to 17.3 percent. Sentence context aids vowel identification by allowing adjustment primarily to a talker's tempo, rather than to the talker's vocal tract.

INTRODUCTION

The acoustic structure of speech varies markedly from one talker to another. The spectrographic measurements of Peterson and Barney (1952) showed that center frequencies of vowel formants vary widely across men, women, and children, and that considerable variation also exists among talkers of the same sex and age group. Similar results were found by Peterson (1961). This acoustic variation is attributed to differences in the sizes and shapes of talkers' vocal cavities. Since each talker's vowels are idiosyncratic in their acoustic composition, it has been thought that a listener needs an extended sample of a talker's speech in order to identify vowel tokens accurately. In general terms, such experience would enable listeners to adjust to each voice they encounter.

Instead of supplying typical frequency values for each vowel, experience with a voice is thought to result in a more general adjustment to the talker's "vowel space." This assumes that a listener identifies a particular vowel of a given talker in terms of the relation between its acoustic structure and the acoustic structure of other vowels produced by the same person (Joos, 1948; Ladefoged and Broadbent, 1957; Ladefoged, 1967). The first sample of a talker's speech will calibrate (or "normalize") the framework to which the listener refers later vowel tokens for identification. Ladefoged and Broadbent (1957) tested this idea with synthetically produced stimuli and found that the perception of an acoustically fixed test word varied predictably as the formant frequencies of a carrier sentence were shifted up or down. They interpreted this result within the framework of adaptation level theory (Helson, 1948), which assumes that perceivers regularly gauge the range of a stimulus continuum in the process of formulating psychophysical judgments.

There have been few explicit hypotheses about how much precursory speech from a talker is required for accurate calibration and what phonetic information is most effective. The most common suggestion, dating back to Joos (1948), is that the point vowels /i, ɑ, u/ are the primary calibrators of vowel space. The most recent proponents of this view are Lieberman and his colleagues (Lieberman, Crelin, and Klatt, 1972; Lieberman, 1973). They argue that experience with the point vowels (or the related glides /j, w/) is a necessary condition for accurate identification of syllables produced by a novel talker. They note that the point vowels are exceptional in several ways: (1) they represent the extreme positions in a talker's articulatory vowel space, (2) they represent the extremes of formant frequency values in a talker's acoustic vowel space, (3) they are acoustically stable for small changes in articulation (Stevens, 1972), and (4) they are the only vowels in which an acoustic pattern can be related to a unique vocal-tract area function (Lindblom and Sundberg, 1969; Stevens, 1972). Other vowels are ambiguous unless calibration to a vocal tract has taken place.

There is little evidence to support the claim of a special role for the point vowels. Suggestive evidence is provided by Gerstman (1968), who developed a computer algorithm for recognition of vowels. Gerstman's algorithm used the extreme values of a talker's formant frequencies (usually those of /i, ɑ, u/) to scale all of the talker's vowels. The algorithm operated on these normalized values and classified the vowels produced by the Peterson and Barney (1952)

panel with a high level of accuracy. However, it must be recognized that such an algorithm is not a perceptual strategy, but only a logically possible strategy. There is no evidence that human listeners perform the computations found in Gerstman's algorithm (such as scaling formants or computing their sums and differences). The results of Ladefoged and Broadbent (1957) provide no assistance on the question of point vowels, since their study did not systematically vary the phonetic content of the precursory speech.

More generally, there is reason to doubt whether a preliminary normalization step plays the major role in vowel perception that is commonly attributed to it. Remarkably low error rates have been found when human listeners identify single syllables produced by human talkers. Peterson and Barney (1952) and Abramson and Cooper (1959) found average error rates of 4 to 6 percent when listeners identified the vowels in h-vowel-d words spoken in random order by a group of talkers. The test words were spoken as isolated syllables, and in most conditions the listeners had little or no prior experience with the talker's voices. On the face of it, these low observed error rates seem inconsistent with any theory that stresses the need for extended prior experience with a talker's vowel space. However, it is difficult to assess the full significance of these findings, since several vowels were substantially more ambiguous than the mean error rates would suggest, and the possible role of point vowels in reducing those ambiguities was not explored.

For these reasons, it is worth investigating what information listeners actually rely upon in natural speech for identifying the vowels produced by a variety of talkers. There is currently no consensus about the perceptual problem posed by vowels in the context of a single syllable, nor about the information gained during experience with a voice. In particular, there is no perceptual evidence that the point vowels play a special role as calibrators of a talker's vowel space. The experiments reported here represent a systematic investigation of these questions.

EXPERIMENT I: PERCEPTION OF VOWELS IN /h-d/ ENVIRONMENT

Identifying a vowel in a naturally spoken syllable should be most difficult when a listener has had no prior experience with the talker's voice. Thus, the need for normalization over several syllables can best be assessed by presenting listeners with a series of single syllables, each spoken by a different talker. The presence of many natural sources of talker-related acoustic variation (for example, differences in age, sex, vocal-tract size, and characteristic pitch level) should maximize the difficulty of such a test. These test conditions were approximated in the perceptual experiments of Peterson and Barney (1952), who presented 20 tokens from each of 10 talkers (men, women, and children) in each block of trials, and Abramson and Cooper (1959), who used 15 tokens spoken by each of 8 adult talkers. Both experiments studied vowels in a fixed /h-d/ consonantal frame.

Our first experiment also used /h-d/ syllables and addressed two major issues: (1) the need for extended familiarization with a talker's vowel space, and (2) the possible role of the point vowels as calibrators of that space. Compared to earlier studies, a greater effort was made in this study to eliminate any potential contribution of familiarity with individual talkers' voices. Thirty talkers each spoke only three syllables distributed throughout the test. In addition, five diphthongs were added to the ten vowels studied by Peterson

and Barney in order to make all perceptual alternatives available to the listeners: /i, ɪ, ɛ, æ, ɑ, ɔ, ʌ, ʊ, u, ɝ, eɪ, oʊ, aɪ, aʊ, ɔɪ/.

There were two test conditions in the experiment. The No-Precursor test contained a long series of /h-d/ syllables; vowel identity and talker identity were unpredictable from one syllable to the next. In the Point-Vowel Precursor test, each /h-d/ test syllable was preceded by a string of three syllables containing the point vowels /i, ɑ, u/ spoken by the same talker. The three vowels were spoken in a /k-p/ consonantal environment; thus, the precursor string contained real words that were different from the test words. The listeners' task in each condition was to identify the vowels in the test syllables. A comparison of the errors made in the two conditions provides a direct measure of the information supplied by exposure to a talker's point vowels. If the point vowels serve as primary calibrators of vowel space, one would expect significantly better vowel identification in the Point-Vowel Precursor condition than in the No-Precursor condition.

Method

1. Stimulus materials. Thirty talkers of varying ages, physical sizes, and characteristic pitch ranges were selected. The group included 13 men, 12 women, and 5 children. All talkers spoke English as their native language, but they were heterogeneous in dialect.

The talkers were recorded individually in a sound-attenuated experimental room with a ReVox A77 stereo tape recorder and Spher-o-dyne microphone. Each talker recorded the full list of 15 test syllables twice, plus two repetitions of the precursor string. The syllables in each precursor string were read at a rate of one per second. The first utterance of each syllable or precursor string was used in the listening tests, unless the talker had clearly mispronounced it.

The test series for each condition contained 90 test syllables, presented in three blocks of 30 syllables each. Each talker contributed only three syllables containing different vowels to the test, one syllable to each block. Each vowel appeared a total of six times, twice within each block. Vowels were assigned to talkers randomly. The order of presentation of syllables within blocks was random, with the following constraints: (1) no less than ten trials intervened between tokens produced by the same talker in one block and the next, and (2) no vowel appeared more than twice in succession.

The Point-Vowel Precursor test was constructed first. Test trials were assembled in the order just described. For each trial, a precursor string was rerecorded, followed by the appropriate test syllable for the same talker. A 1-sec pause was inserted between the last precursor syllable and the test syllable. The same precursor string preceded all three of a talker's test syllables. Peak intensity for each precursor string and test syllable was equalized within 0.5 dB as monitored on the VU-meter of the tape recorder. A 4-sec intertrial interval was inserted between each test syllable and the following set of precursors, and a 10-sec interval was inserted between blocks of 30 syllables.

The No-Precursor test was constructed by rerecording the test syllables and deleting the precursors. Thus, the two tests contained identical test syllables; the order of presentation, the intervals between successive test syllables, and the intensity of the syllables were all the same.

2. Procedure. Tests were presented to small groups of subjects in a quiet experimental room via a Crown CX 822 tape recorder, MacIntosh MC40 amplifier, and AR acoustic suspension loudspeaker. The output level was the same for both tests, as monitored by a Heathkit AC VTVM placed just ahead of the output to the loudspeaker. The level was clearly audible in all parts of the room. Subjects responded on score sheets that contained 15 response alternatives, all written out in full and arrayed in rows as follows: "hood, head, hoed, heard, who'd, hide, heed, how'd, hud, hayed, hod, hoyed, had, hid, howed." They were told that they would hear "several different talkers." Subjects in the Point-Vowel Precursor condition were informed that each test word would be preceded by three other words spoken by the same person, and that listening to those three words might help them identify the fourth. Subjects listened to the full test series twice, for a total of 180 judgments per subject, 12 on each intended vowel.

3. Subjects. The listeners were 37 paid volunteers from undergraduate psychology classes at the University of Minnesota. All were native speakers of English and most were native to the upper midwest region of the United States. Seventeen were subjects in the No-Precursor condition, while 20 were subjects in the Point-Vowel Precursor condition.

Results and Discussion

Errors in vowel identification were tabulated for each condition. An error was defined as a failure to select the vowel intended by the talker: the error category included omissions, that is, failures to select any alternative. In the No-Precursor condition, subjects made an average of 12.9 percent errors, and in the Point-Vowel Precursor condition, subjects averaged 12.2 percent errors on the test syllables. Contrary to the prediction that point-vowel precursors would substantially reduce errors, the error rates for the two conditions were not significantly different [$t(35) = 0.57$].

The error rate in the No-Precursor condition was somewhat higher than the error rates found in the two earlier studies using /h-d/ syllables. Peterson and Barney (1952) reported an overall error rate of 5.6 percent. Their lower observed rate may be due to the smaller number of response alternatives in their study (10 instead of 15), the smaller number of talkers appearing in a particular block of trials (10 instead of 30), and the larger total number of tokens from each talker (20 instead of 6). Abramson and Cooper (1959) reported an error rate of 4.0 percent in a study involving 15 vowel alternatives and eight adult talkers. In contrast to the present study, talkers carefully selected tokens they considered typical, and the listeners were familiar with the talkers (in fact, the group of listeners included the talkers). In addition, the number of talkers in the Abramson and Cooper study was smaller (8 instead of 30) and the total number of tokens from each talker was larger (15 instead of 6). Thus there are several possible sources for the higher error rate observed in the No-Precursor condition of this study. But whatever the source, it must not be overlooked that 12.9 percent is a remarkably low error rate for a 15-alternative response set, especially if one believes that a single syllable from a novel talker is a highly ambiguous entity.

Though experience with talkers' point vowels did not reduce overall errors, it is important to determine whether the precursors influenced the perception of individual vowels. The percentage of errors made on each intended vowel is presented in Table 1 for each test condition. (Confusion matrices for these

conditions are presented in Tables A-1 and A-2 in the Appendix.) Several results are worth noting. First, errors tended to be very high on the intended vowels /ɑ/ and /ɔ/. Most of these errors involved confusions between the two vowels. In fact, confusions between /ɑ/ and /ɔ/ account for 39 percent of all errors made by listeners in the No-Precursor condition, compared to 28 percent of all errors in Peterson and Barney's (1952) experiment. Thus, the phonetic confusion between /ɑ/ and /ɔ/ may have contributed to the higher overall error rate observed in this study. The degree of confusability is not surprising since little distinction is made between /ɑ/ and /ɔ/ in upper midwestern dialects; most of the listeners (and many of the talkers) were native to that region. The error rates for identifying these two vowels, excluding /ɑ/-/ɔ/ confusions, are included in parentheses in Table 1.

TABLE 1: Mean percent error in identification of /h-d/ syllables.

	Condition	
Intended vowel	No-precursor	Point-vowel precursor
i	1.0	0.0
ɪ	20.1	29.6
ɛ	19.1	9.2
æ	12.3	9.6
ɑ	48.5 (9.3)[a]	43.3 (4.6)
ɔ	18.1 (9.3)	42.9 (19.2)
ʌ	14.7	3.8
ʊ	14.7	18.3
u	8.3	1.7
ɝ	0.0	0.0
eɪ	2.4	2.1
oʊ	12.7	4.6
aɪ	2.0	0.0
aʊ	16.2	17.9
ɔɪ	3.9	0.0
Overall	12.9 (9.7)	12.2 (8.0)

[a]Parenthesized figures present the mean percent error when confusions between /ɑ/ and /ɔ/ are excluded.

Second, several vowels were identified very accurately, even in the No-Precursor condition. This is true for two of the three point vowels (/i/ and /u/), for /ɝ/, and for three of the diphthongs (/eɪ/, /aɪ/, and /ɔɪ/). Low error rates for /i/, /u/, and /ɝ/ were also observed by Peterson and Barney (1952). The presence of two point vowels in this group verifies predictions that they should be relatively unambiguous (cf. Lieberman et al., 1972), although their role as calibrators remains in question. The low error rates for diphthongs suggests that their addition to the response set did not contribute much to the higher overall error rate in this study. The error rate for the five diphthongs averaged only 6 percent across the two conditions.

Third, and most importantly, there was no consistent pattern of change when test syllables were preceded by point-vowel precursors. This was true even for the relatively ambiguous vowels. Of the seven vowels showing a greater-than-average number of errors in the No-Precursor condition, three showed an apparent improvement following precursors (/ɛ/, /ɑ/, /ʌ/), while four showed an increase in errors (/ɪ/, /ɔ/, /ʊ/, /aʊ/). Thus, in terms of overall errors on individual vowels, there was no consistent support for the hypothesis that experience with a talker's point vowels allows a listener to disambiguate troublesome vowels.

The differences in error rate for individual vowels need to be interpreted with caution. Differences in response biases in the two conditions could have been responsible for some of the apparent changes in identifiability. That is, a vowel could have been correctly identified more often simply because it was more popular as a response. One indication of such a response bias is how often a vowel is used as an incorrect response to other vowels; when the vowel becomes more popular, the frequency of these false identifications increases. Figure 1 depicts the results of a preliminary analysis for response biases. The horizontal axis indicates the change in correct identification (in percent) between the Point-Vowel Precursor and No-Precursor conditions. Placement to the right of the central vertical line represents superior performance in the Point-Vowel Precursor condition compared to that in the No-Precursor condition. The vertical axis indicates the change in false identification. (This is defined as the percentage of vowel tokens incorrectly identified as a particular vowel.) Placement above the central horizontal line represents a greater frequency of false identifications in the Point-Vowel Precursor condition relative to the No-Precursor condition.

In this preliminary analysis, "true" improvements attributable to precursors may be defined by an increase in correct responses, coupled with a decrease in false identifications.[1] Of the vowels that were most ambiguous in the No-Precursor condition, only /ɑ/ showed genuine improvement by this measure. Several less ambiguous vowels also showed genuine improvement: /æ, u, oʊ, aɪ, ɔɪ/. On the other hand, a change in correct identification that corresponds in sign with a change in false identification may be referred to descriptively as a

---

[1] It is important to note that the relationship between the scales on the horizontal and vertical axes is arbitrary. For example, if a vowel appears in the upper right-hand quadrant on a 45° line passing through the origin, this cannot be interpreted as an increase in correct responding that is "perfectly correlated" with the increase in false responding. In Figures 1, 2, and 3, the aspect ratios have been chosen so that the ranges of values on each dimension are given roughly equal weight. It is also important to note that the differences plotted are linear functions of error scores. On either axis, the differences indicate the relative contribution of each vowel to the overall change in percent identification. However, the values plotted give no indication of the proportionate change in identification on each vowel. For example, if vowel x increased in correct identification from 50 to 55 percent, and vowel y increased from 94 to 99 percent, each would appear along the horizontal axis at +5 percent, though the proportionate improvement is larger for y. The primary goal of these figures is their heuristic value in visualizing relative directions of change in two variables. Choice of the linear transform should not be interpreted as a claim about what differences represent "equivalent" changes in the recognition system.

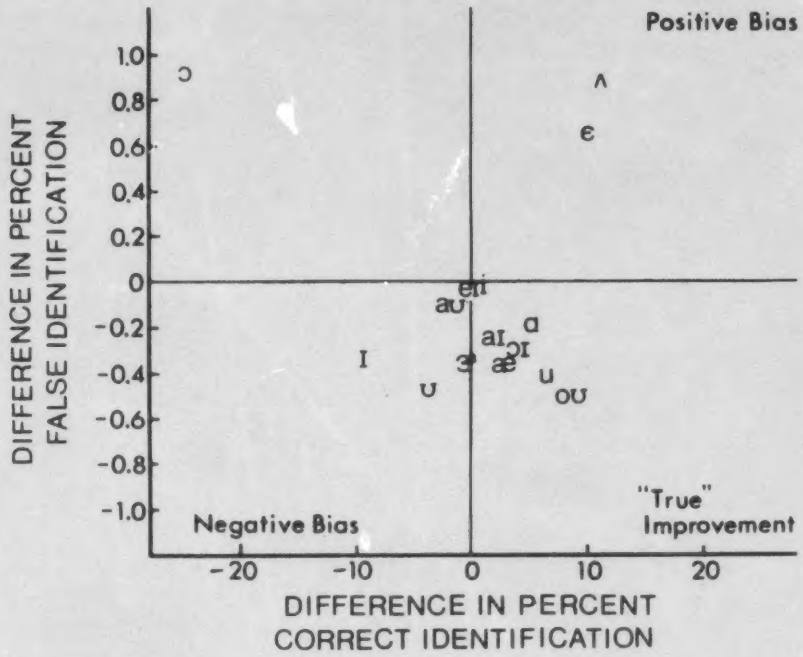


Figure 1: Changes in correct and false identification attributable to /kip, kɑp, kup/ precursors (/h-d/ syllables). Each axis plots the difference between the Point-Vowel Precursor condition and the No-Precursor condition.

"positive" or "negative bias." Two vowels, /ɛ/ and /ʌ/, showed a clear positive bias, while /ɪ/, /ʊ/, and /aʊ/ showed a negative bias. The remaining ambiguous vowel /ɔ/ showed no sign of improvement: a large increase in false responses was associated with a large decrease in correct responses.

The analysis displayed in Figure 1 cannot indicate which changes are significant departures from chance variability, nor can it fully disentangle changes in stimulus identifiability from changes in response biases. The number of false identifications of a vowel x might increase, not because of an increased response bias toward x, but because the perceptual similarity (confusability) of x with another vowel y may have increased. Correct and false identification scores for x will reflect the combined impact of changes in the similarity of x to several other vowels (some similarities may increase, while others decrease) and changes in response biases of all vowels concerned. Luce's Choice Axiom (Luce, 1959, 1963) provides one means of modeling these interactions in a confusion matrix. The model assigns a similarity parameter η_{xy} to each pairwise combination of stimuli and a response bias parameter β_y to each response alternative. The combined action of these parameters determines a predicted distribution of responses in the confusion matrix.

The Luce model is useful because it allows one to assess the significance of changes in a similarity parameter from one condition to another.[2] In the present experiment, any beneficial effect of hearing point-vowel precursors should manifest itself in a decrease in pairwise similarity measures (i.e., pairwise confusions should decrease). Of the 105 possible pairwise combinations of 15 stimuli, 12 pairs accounted for 81 percent of the errors in the No-Precursor condition and 88 percent of the errors in the Point-Vowel Precursor condition. Similarity measures were determined for each of these pairs, and a t-statistic was computed to assess the significance of the difference between the measures for the two conditions. Only two of the pairs showed a significant change in similarity following point-vowel precursors: /ɑ-ɔ/ and /ɔ-aʊ/; both were cases of <u>increased</u> confusability and both involved the vowel /ɔ/. This was a genuine decrement in performance on /ɔ/, which cannot be attributed to an overall change in response biases (as might be expected from Figure 1). None of the other confusable pairs showed significant changes in similarity.

These results have direct implications for the six vowels in Figure 1 that showed change in the direction of "true" improvement: /æ, ɑ, u, oʊ, aɪ, ɔɪ/. The confusion pairs for which similarity measures were obtained include the major sources of error for each of these vowels. With one exception, none of these sources of error showed a significant effect of point-vowel precursors. The exception was the confusability of /ɑ/ and /ɔ/, which showed a large increase. (The increase appeared mainly in incorrect /ɔ/ responses to /ɑ/, possibly due to a contrast between tokens of /ɑ/ in the precursor strings and the test syllables.) In general, then, even the "true" improvements cannot be interpreted as anything more than expressions of chance variability.

Thus, the patterns of error with and without point-vowel precursors were similar, showing major differences only in the identification of /ɔ/. The presence of these differences indicates that the precursors did have an impact on subjects' judgments; the nonsignificant difference in overall errors between the two conditions cannot be due to inattention to the precursor strings. Even so,

---

[2] The predicted frequency of identifying an intended vowel $\underline{x}$ as the response alternative $\underline{y}$, e_{xy}, is defined by the formula:

$$e_{xy} = \frac{\beta_y \eta_{xy} n_x}{\sum_{y=1}^{N} \beta_y \eta_{xy}}$$

where N is the number of vowel categories (15 in Experiment I) and n_x is the total number of intended vowels that were presented (12 per subject in Experiment I). These "expected values" were estimated for each cell of the confusion matrices, using an algorithm developed by J. E. Keith Smith at the University of Michigan. At theoretical limit, the procedure outputs the set of maximum likelihood estimators for the observed pattern of errors. The x-y similarity parameters were estimated as follows: $\eta_{xy} = (e_{xy}e_{yx}/e_{xx}e_{yy})^{1/2}$. Since $-\ln \eta_{xy}$ closely approximates a normal distribution, similarity parameters for two conditions may be compared using the t-statistic, $t = 2(\ln \eta_2 - \ln \eta_1)/(V_1 + V_2)^{1/2}$, where V is the estimated variance. A full development of this general procedure may be found in Goodman (1969, 1970).

there is no support in these results for the point-vowel hypothesis; the major differences involved increases in ambiguity and shifts in response biases.

Perhaps the most striking result is that subjects generally had little difficulty identifying the test syllables, even when there was no prior information about talkers' vocal tracts. It is possible that the level of identification was so high in the No-Precursor condition that there was little room for improvement: 87 percent may represent a ceiling on identifiability of these test syllables under any conditions. Thus the failure to find a precursor effect in this experiment might indicate (1) that point vowels do not bear the kind of information hypothesized, or (2) that there may be no need for such information, if there are no errors that are a function of uncertainties in normalization. It is necessary to know what component (if any) of the 12.9 percent error rate is due to subjects' uncertainty about the vocal tracts to which they are listening. This would define the maximum improvement in identification that could be contributed by the presence of precursors. The next experiment was designed to measure the error component attributable to vocal-tract uncertainty and to reassess the potential value of sample vowels in reducing that uncertainty.

EXPERIMENT II: THE PERCEPTION OF VOWELS IN /p-p/ ENVIRONMENT

Two conditions in this experiment were designed to measure the error component in vowel perception that is attributable to talker variation. In the Mixed Talker condition a large number of talkers spoke a series of syllables; on each test syllable the listener encountered a voice that was unfamiliar and unpredictable. (This condition is comparable to the No-Precursor condition of Experiment I.) In the Segregated Talker condition subjects heard the same series of syllables spoken by one person, so there was ample opportunity to become familiar with the voice and the talker was fully predictable from one syllable to the next. The difference between the error rates in these two conditions provides a measure of the increment in perceptual error introduced by talker variation.

Two additional mixed talker conditions were included to reassess the role of precursory information in reducing perceptual errors. In each condition, the test syllables of the Mixed Talker test were preceded by a precursor string from the appropriate talker. In the Point-Vowel Precursor condition, the precursor string was /hi, hɑ, hu/ (/h-/ syllables were chosen to facilitate articulation, while minimizing nonvocalic sources of information). In the Central-Vowel Precursor condition, each syllable was preceded by /hɪ, hæ, hʌ/.[3] As was argued in Experiment I, point-vowel precursors should substantially reduce errors if they are privileged carriers of information for normalization. A comparable set of nonpoint vowels should produce little or no improvement in identification, by the same hypothesis. Finally, if the information available in point vowels is essentially that gained during extended familiarization with a vocal tract, then performance in the Point-Vowel Precursor condition should resemble that in the Segregated Talker condition.

---

[3] The term "central vowel" is used only in contrast to "point vowel," not in the more restricted sense found in traditional phonetic taxonomies. Of the six central vowels so defined, a set of three with fairly wide dispersion in two-formant space were chosen for this condition.

Several changes made in the design of this experiment were intended to increase the average level of errors beyond that found in Experiment I. First, the consonantal context for the vowels was changed from /h-d/ to /p-p/. The /p-p/ environment was chosen because vowel duration tends to be shorter in voiceless stop contexts than in voiced contexts (Stevens and House, 1963). Second, an effort was made to reduce syllable duration and increase coarticulation effects by encouraging talkers to speak rapidly when recording the syllables. Third, the five diphthongs and /ɝ/ were eliminated from the vowel set, since they tended to produce few errors and would be relatively uninformative in the present design.

Method

1. Stimulus materials. A panel of 15 talkers (five men, five women, and five children) was chosen to produce the test syllables for the mixed talker conditions. They were selected to represent a wide variety of vocal-tract sizes and characteristic fundamental frequencies. None were phonetically trained speakers. In the judgment of the experimenters, the talkers represented a fairly homogeneous dialect group, that of the upper midwest region from which the listeners were also drawn.

The Mixed Talker tests consisted of 45 tokens, 5 tokens of each of the 9 syllables: /pip/, /pɪp/, /pɛp/, /pæp/, /pɑp/, /pɔp/, /pʌp/, /pʊp/, and /pup/. Each talker contributed three test syllables. Vowels were randomly assigned to talkers with the constraint that each talker contributed three different vowels, only one of which was a point vowel (/i/, /ɑ/, or /u/). Thus, the five tokens of each syllable type were spoken by different talkers. In addition to three test syllables, each talker produced two sets of precursors: /hi, hɑ, hu/ and /hɪ, hæ, hʌ/. The syllables in each triplet were read at a rate of one per second. No attempt was made to control the intonation pattern of the three-syllable utterance.

The 45 recorded syllables for the Mixed Talker test were arranged in a random presentation order with the constraints that (1) the same intended vowel did not appear more than twice consecutively, and (2) tokens produced by the same talker were separated by not less than 8 tokens. A 4-sec interval was inserted between tokens, and a 10-sec interval was inserted after each block of 15 tokens.

The Point-Vowel Precursor test was constructed by inserting copies of each talker's point-vowel triplet in front of the appropriate three test syllables in a copy of the Mixed Talker test. In each case a 1-sec interval was inserted between the offset of the final precursor syllable and the test syllable.

The Central-Vowel Precursor test was constructed using each talker's central-vowel triplet, according to the same procedures. Thus, all three Mixed Talker tests contained identical test syllables; the order of presentation, the intensity levels, and the intertrial intervals were all the same.

For the Segregated Talker test, one representative man, one woman, and one child were selected from the full panel of talkers.[4] For each component test (Man,

[4]The man, woman, and child chosen as "representative" were individuals in each group of talkers whose test syllables produced a close-to-average number of

Woman, Child) the talker produced the full series of 45 test syllables, five different tokens of each of the nine syllable types. The 45 tokens were arranged in the same order as in the Mixed Talker test.[5]

2. Procedure. Tests were presented to small groups of subjects under the same listening conditions as in Experiment I. Subjects responded on score sheets that contained nine response alternatives in each row: "pip, pup, pap, peep, pop, pep, poop, pawp, puup." The experimenter pronounced each word, drawing special attention to the last word, "puup," which stood for the syllable /pup/. The three Mixed Talker tests were presented to independent groups of subjects. Subjects completed two repetitions of the 45 test trials, for a total of 90 judgments per subject, 10 on each intended vowel. Three additional groups of subjects listened to the Segregated Talker tests; each group completed all three tests: Man (M), Woman (W), and Child (C). The order of presentation of the tests was counterbalanced across groups in the orders: MWC, WCM, and CMW. For each group of subjects, data from only the first two tests were analyzed. Thus the total number of judgments for the Segregated Talker condition was equal to that for each Mixed Talker condition (90 judgments per subject) and any effects of fatigue or task familiarity were equally distributed across the three talkers in the Segregated Talker tests.

3. Subjects. The listeners were 79 paid volunteers from undergraduate psychology classes at the University of Minnesota. All were native speakers of English and most were native to the upper midwest region of the United States. In mixed talker conditions, 19 subjects heard the Mixed Talker test, 15 heard the Point-Vowel Precursor test, and 12 heard the Central-Vowel Precursor test. The remaining 33 subjects served in the Segretated Talker condition; 11 subjects heard each of the counterbalanced orders.

Results and Discussion

In the Mixed Talker condition (without precursors), subjects made an average of 17.0 percent errors in identifying vowels produced by the panel of randomly ordered talkers, while in the Segregated Talker condition, listeners averaged 9.5 percent errors for the vowels of the three single talkers. [The mean error rates for the individual tests were 9.8 percent (Man), 6.8 percent (Woman), and 11.8 percent (Child).] Familiarity with a talker's voice significantly improved the accuracy of identification [$t(50) = 5.14$, $p < .01$]. Even so, this factor accounts for less than half of the errors in the Mixed Talker condition.

There are two ways to look at the error percentages for /p-p/ syllables. First, on the Segregated Talker test, 9.5 percent is a relatively high error rate, considering the complete predictability from trial to trial to both the talker's

---

errors on the Mixed Talker test, and who were available for further recording sessions.

[5] Acoustic measurements of vowels in the Mixed and Segregated Talker tests are reported in a companion study (Strange, Verbrugge, Shankweiler, and Edman, in press). Average formant frequency and relative duration values were comparable to those reported by Peterson and Barney (1952), Peterson and Lehiste (1960), and Stevens and House (1963).

voice and the consonantal frame. There are sources of vowel ambiguity not attributable to uncertainties in calibration. Second, on the Mixed Talker test, 17 percent is a relatively low error rate, given that each judgment is made with no familiarity with the voice and without the benefit of sentence context. This error rate is not substantially greater than the overall 12.9 percent rate found for /h-d/ syllables in a similar mixed talker test (No-Precursor condition, Experiment I), though several changes were made that were intended to increase errors.[6] There is clearly a great deal of information within a single syllable that specifies the identity of its vowel nucleus.

The data for the Mixed and Segregated Talker conditions challenge the assumption that extended familiarization with a vowel space is the primary factor controlling vowel identification. Even so, some information must be available in a series of utterances from a single talker, since listeners correctly identified more vowels in the Segregated Talker test than in the Mixed Talker test. A vowel-by-vowel analysis of subjects' errors indicates that this improvement was not distributed evenly among the nine vowels. The first two columns in Table 2 present the error rate for each intended vowel in the Mixed and Segregated Talker conditions. Three of the vowels /i, ɪ, u/ showed little change, since almost all tokens were correctly identified in both conditions. Of the six relatively ambiguous vowels, only /ɑ/ failed to show improvement, while familiarization aided perception of /ɛ, æ, ɔ, ʌ, ʊ/. (Confusion matrices for these two conditions are presented in Tables A-3 and A-4.)

TABLE 2: Mean percent error in identification of citation-form /p-p/ syllables.

Intended vowel	Condition: Mixed talker	Segregated talker	Point-vowel precursor	Central-vowel precursor
i	1.1	0.3	3.3	3.3
ɪ	1.6	3.6	2.7	1.7
ɛ	26.8	12.1	4.7	10.8
æ	18.9	1.8	20.7	18.3
ɑ	20.0 (10.0)	22.7 (3.9)	43.3 (26.7)	29.2 (12.5)
ɔ	27.4 (3.2)	18.5 (1.8)	18.7 (12.7)	13.3 (2.5)
ʌ	15.3	7.6	9.3	22.5
ʊ	38.9	17.6	26.7	29.2
u	2.6	0.9	7.3	5.8
Overall	17.0 (13.2)	9.5 (5.5)	15.2 (12.7)	14.9 (11.9)

As in Experiment I, it is important to isolate the contribution of response biases and to discover whether any of the changes in vowel similarity reflect factors other than chance variation. Again, both a graphic analysis and the

[6] The shift to a /p-p/ consonantal frame apparently had little effect on the error rate for the nine vowels studied here. Errors on those nine vowels averaged 17.4 percent in /h-d/ syllables (with 15 response alternatives), compared to 17.0 percent in /p-p/ syllables.

Luce choice model were applied to the data from the Segregated and Mixed Talker conditions. The first analysis (presented in Figure 2) showed "true improvement" in the identification of /ɛ/, /æ/, /ʌ/, /ʊ/, and /u/ in the Segregated Talker condition. The apparent improvement for /ɔ/ was associated with a large positive bias, while /ɑ/ showed a negative bias. The Luce similarity analysis showed significantly reduced confusions between the following pairs: /ɛ-æ/, /ɑ-ʌ/, /ʌ-ʊ/, and /ʊ-u/. These four confusable pairs were major sources of error for the five vowels showing true improvement. Thus, the increases in correct identification for these vowels reflect more than chance variation. They represent genuine compensation for confusions due to talker variation.

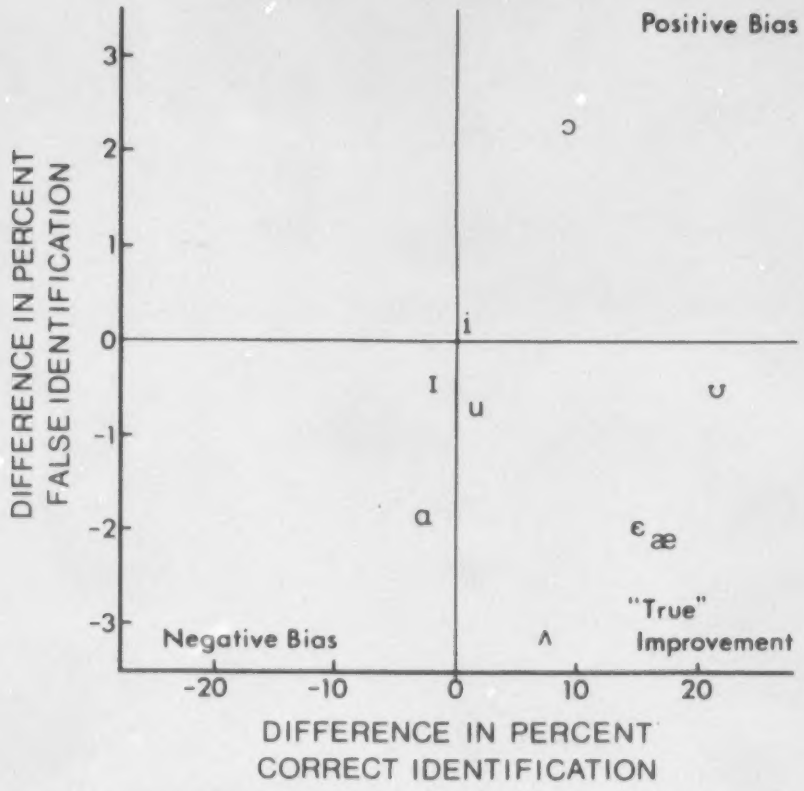


Figure 2: Changes in correct and false identification attributable to keeping the talker constant throughout a test (citation-form /p-p/ syllables). Each axis plots the difference between the Segregated Talker condition and the Mixed Talker condition.

The failure to find true improvement for either /ɑ/ or /ɔ/ or a significant decrease in their pairwise confusion reflects their somewhat ambiguous status in upper midwestern dialects. On the average, errors for /ɑ/ and /ɔ/ were almost as frequent for a single talker as they were for a mixed group of talkers. Thus, the similarity of /ɑ/ and /ɔ/ is apparently a function of the dialect, not of unfamiliarity with talkers' voices.

The kind of improvement resulting from familiarization with a talker's vowel space may be summarized as follows: overall errors drop somewhat (7.5

percent in this experiment), genuine overall improvement is found for several ambiguous vowels, and there is a significant decrease in similarity for several vowel pairs. If the point vowels specify efficiently the kind of information gained during extended familiarization, we would expect a similar pattern of improvement in the Point-Vowel Precursor condition.

The results did not support this hypothesis. Exposure to a talker's point vowels aided listeners only slightly, reducing overall errors from 17.0 to 15.2 percent; the difference was not statistically significant [$t(32) = 0.97$]. In the Central-Vowel Precursor condition, overall errors also dropped slightly, to 14.9 percent, though again the change was not significant [$t(29) = 1.21$]. In other words, not only was there no evidence for a gain attributable to point vowels, but there was no difference between the point vowels and a set of non-point vowels. In general, experience with specific sets of vowels seems to make little contribution to the total reduction of errors attributable to prior experience with a person's voice.

It is important to determine whether these conclusions are affected by the results for individual vowels. The right-hand columns in Table 2 present the errors on each intended vowel following point-vowel and central-vowel precursors. (Confusion matrices for these conditions are presented in Tables A-5 and A-6.) A comparison of errors in the Point-Vowel Precursor condition and the Mixed Talker condition (without precursors) is presented in Figure 3. In general, the point vowels did not produce a "true improvement" in the perception of ambiguous vowels like that found in the Segregated Talker condition. Where similar apparent improvements were found, they tended to be associated with much higher relative levels of false identification in the Point-Vowel Precursor condition (compare Figures 2 and 3). In other cases, apparent improvements found for the Segregated Talker condition were not found with the point-vowel precursors. A Luce analysis indicated that the only comparable change in pairwise similarities was a substantial reduction in /ɛ-æ/ confusions in both conditions. None of the other reductions found with segregated talkers were found with point-vowel precursors. In addition, the /ɔ-ʌ/ confusion, which showed no change with segregated talkers, showed a sharp increase in the Point-Vowel Precursor condition.

When the Central-Vowel Precursor condition was compared to the Mixed Talker condition on a vowel-by-vowel basis, virtually the same results were obtained. No vowel showed more than a marginal change in the direction of true improvement, and a significant decrease in pairwise similarity was observed for /ɛ-æ/. However, the increase in the /ɔ-ʌ/ confusion observed with point-vowel precursors was not observed here. Thus, to the limited extent that improvements are found at all with precursors, there is no evidence that the three point vowels are unique as sources of information about a talker's vowel space.

In general, however, neither set of vowel precursors were efficient carriers of the kind of information available in extended experience with a talker's voice. Sets of vowels of known identity did not produce reductions in overall errors, errors on specific vowels, or pairwise similarities comparable to those produced by extended experience.

An extension of the Luce model allows one to make comparisons between the overall error patterns for two experimental conditions. Specifically, one may ask whether the same set of stimulus similarity and response bias parameters is sufficient to describe both patterns, or whether different sets provide a closer

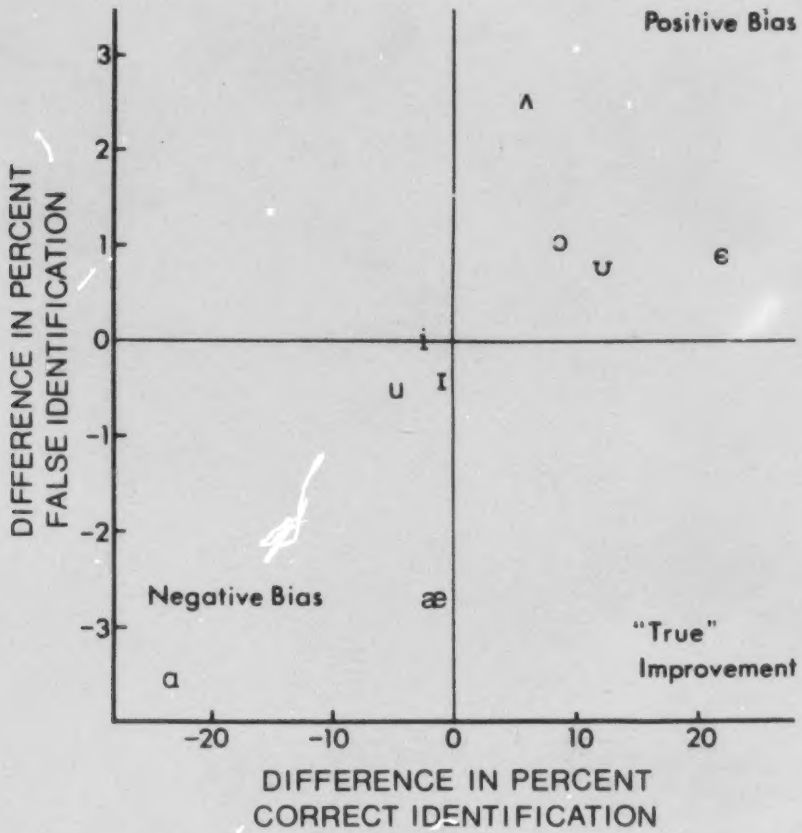


Figure 3: Changes in correct and false identification attributable to /hi, hɑ, hu/ precursors (citation-form /p-p/ syllables). Each axis plots the difference between the Point-Vowel Precursor condition and the Mixed Talker condition.

fit. In the latter case, one may test models in which only the similarity parameters for each condition differ, in which only the bias parameters are different, or in which both parameter sets differ.

Joint models for the Mixed and Segregated Talker conditions suggest that the dominant impact of extended familiarization is on perceptual similarity. The different-parameters model ($\chi^2/df = 3.54$), in which both sets differ, provides a closer fit than the same-parameters model ($\chi^2/df = 5.43$).[7] This improvement is

[7] For ease of comparison, the goodness-of-fit for each model has been characterized by the ratio of the maximum-likelihood χ^2 value to the number of degrees of freedom. Most of the χ^2 values are significant, and the Luce models appear to be rejected. However, these significance tests assume that the observed frequencies manifest stable population probabilities. Analysis of the variability among subjects revealed significant heterogeneity in their responses to several vowel categories. Thus, the reported χ^2 values reflect substantial heterogeneity among subjects, as well as deviations of the expected values from underlying population values. When adjustments are made for the observed heterogeneity, the fit of the Luce models is much improved. The unadjusted χ^2 ratios provide a useful measure for present purposes, since the degree of heterogeneity was roughly constant across the experimental conditions being compared.

contributed largely by different similarity parameters: the different-similarities model ($\chi^2/df = 3.83$) fits both conditions more sucessfully than the different-biases model ($\chi^2/df = 5.27$). This means that the main effect of hearing a single talker is on a listener's ability to discriminate the vowels themselves, not on the listener's response biases.

A different result is found when the Mixed Talker condition is compared with each precursor condition. In each case, estimating different similarity parameters fails to improve the overall goodness-of-fit; different bias parameters, on the other hand, do improve the model. When errors in the Mixed Talker and Point-Vowel Precursor conditions are jointly modeled, the same-parameters model ($\chi^2/df = 3.68$) fits substantially better than the different-similarities model ($\chi^2/df = 5.78$), but not as well as the different-biases model ($\chi^2/df = 2.46$). Similarly, when errors in the Mixed Talker and Central-Vowel Precursor conditions are jointly modeled, the same-parameters model ($\chi^2/df = 2.66$) is not improved by the addition of different similarity parameters ($\chi^2/df = 4.55$), but is improved by different bias parameters ($\chi^2/df = 2.35$). Thus, the precursors not only produced a pattern of similarity changes different from that hypothesized, but produced change of a different kind altogether. Precursors predominantly affected listeners' preferences for various response alternatives, rather than their ability to distinguish among intended vowels.

A possible shortcoming of the design of this experiment is that the test syllables were not sufficiently "natural": since they were spoken in citation form, the formant frequencies of their vocalic centers would not show the degree of variability found for destressed vowels in rapidly articulated sentences. It is possible that the task of perceiving rapidly spoken syllables places a higher premium on information about the vocal tract. Experiment III was designed to determine whether point vowels would benefit listeners on a mixed talker task involving rapidly articulated vowels.

EXPERIMENT III: PERCEPTION OF VOWELS IN DESTRESSED /p-p/ SYLLABLES

In the rapidly articulated syllables of connected speech, vowel durations tend to be short and vowel formants are not likely to reach steady-state values. Formant values at the center of syllables in connected speech are different from those found in single syllables spoken in citation form, the degree of deviation depending systematically on the rate of articulation and the amount of destressing (Tiffany, 1959; Shearme and Holmes, 1962; Lindblom, 1963; Gay, 1974). If vowel perception involves relating vowels to a "space" (defined by some transformation on formant frequencies), then the frequency variation contributed by speaking rate should considerably enhance a listener's difficulty in calibrating to a talker's space. This experiment explores the perceptual problem posed when both talker-dependent and rate-dependent variation are present. The error rate for single, rapidly articulated syllables excised from carrier sentences should be substantially greater than that found for syllables spoken in isolation. Given the (presumably) more difficult task of identifying a rapid, destressed syllable, information about a talker's point vowels may play a larger role than was found in preceding experiments.

The experiment consisted of three test conditions. In the No-Precursor condition, listeners heard a mixed talker test containing /p-p/ syllables spoken by the same panel of talkers used in Experiment II. The syllables were spoken

in destressed position in the context of a full carrier sentence and were excised for use in the test. In the Point-Vowel Precursor condition, each test syllable was preceded by a point-vowel precursor string spoken by the appropriate talker. In the Sentence Context condition, each test syllable was heard in the context of the carrier sentence in which it was originally produced. One would expect the error rate in this condition to be lower than that in the No-Precursor (and no context) condition, since more information is available about the talkers prior to the test syllables. If so, the degree of improvement provides a measure of the information supplied by sentence context, when no semantic factors are involved. The pattern of improvement following point-vowel precursors should be similar, if the predominant effect of both types of context (precursor and sentence) is to allow calibration to a talker's vowel space.

Method

1. Stimulus materials. Each of the 15 talkers contributed the same three syllables they had produced for the mixed talker tests in Experiment II. In all three conditions of this experiment, the order of talkers and test syllables was the same as in the earlier experiment. The tests contained five tokens of each of nine /p-p/ words; each of the five tokens was produced by a different talker and each talker contributed only one point vowel. The test syllables were spoken in the following carrier sentence: "The little p-p's chair is red." Talkers were instructed to read each sentence rapidly, stressing the word "chair."

The test syllables were excised from copies of the carrier sentences for use in the No-Precursor and Point-Vowel Precursor tests. Each recording was monitored and the audio tape was cut within the silent interval just preceding the release burst of the initial /p/ and during the silent closure interval of the final /p/. Thus, the final /p/ of the test syllables did not include a release from closure. To produce the No-Precursor test, the 45 excised syllables were assembled in the presentation order and then rerecorded as in Experiment II. The Point-Vowel Precursor test was constructed by inserting copies of each talker's point-vowel triplet in front of the appropriate three test syllables in a copy of the No-Precursor test, using the same precursor strings and recording procedure as in Experiment II. Thus, the No-Precursor and Point-Vowel Precursor tests contained identical test syllables, with the same order of presentation, intensity levels, and intertrial intervals, and each was comparable in these respects to the mixed talker conditions of Experiment II. The Sentence Context test was constructed using copies of the original carrier sentences. The order of talkers and component test syllables was the same as that in the other two tests. A 4-sec interval was inserted between each sentence.

2. Procedure. Tests were presented to small groups of subjects under the same conditions as in previous experiments. Subjects in the Sentence Context condition were told that each test word would be spoken in the middle of the same sentence: "The little (something)'s chair is red." The three tests were presented to independent groups of subjects. Subjects completed two repetitions of the 45 test trials, for a total of 90 judgments per subject, 10 on each intended vowel.

3. Subjects. The listeners were 52 paid volunteers from undergraduate psychology classes at the University of Minnesota. All were native speakers of

English and most were native to the upper midwest region. Twenty were subjects in the No-Precursor condition, 17 in the Point-Vowel Precursor condition, and 15 in the Sentence Context condition.

Results and Discussion

Listeners averaged 23.8 percent errors in identifying the vowels in the excised syllables without precursors. As expected, this error rate is higher than the 17.0 percent rate found for citation-form syllables in the comparable Mixed Talker test in Experiment II; the difference between these two conditions is significant [$t(37) = 3.88$, $p < .01$].

Given the increased ambiguity when both talker- and rate-dependent variation are present, it might be expected that listeners would make greater use of a talker's point vowels to reduce that ambiguity. Contrary to this expectation, the average error rate in the Point-Vowel Precursor condition was 28.6 percent, which is significantly higher than the 23.8 percent rate found when no precursors are present [$t(35) = 2.85$, $p < .01$]. This is a startling result: it does not fulfill the expectation that greater improvement would be found where more was needed, nor does it even replicate the minor improvements found with point-vowel precursors in Experiments I and II.

In contrast to these results for point-vowel precursors, a substantial decrease in errors was found when the test syllables were heard in their original sentence context. Listeners made an average of 17.3 percent errors in the Sentence Context condition; this is significantly lower than the 23.8 percent error rate found for the test syllables in excised form [$t(33) = 3.31$, $p < .01$]. Thus, a carrier sentence contains information that makes vowels in component syllables less ambiguous.

Error rates for individual vowels are presented in Table 3 for each of the three test conditions. A comparison of the results for excised syllables (first column, Table 3) and for citation-form syllables (first column, Table 2) suggests that listeners in the No-Precursor condition may not have accommodated completely to the rapid pace at which the excised syllables were spoken. In general, errors on these syllables were in the direction of hearing vowels in the periphery of two-formant space as more "centralized" or "reduced" (cf. confusion matrix, Table A-7). (1) Two point vowels, /i/ and /u/, which produced very few errors in citation-form syllables, were somewhat ambiguous in the destressed syllables. The errors on /i/ generally involved misperceiving it as /ɪ/. The vowel /u/ tended to be misperceived as /ʊ/. (2) Errors more than doubled on /ɑ/ and /ɔ/. By far the most common error on both /ɑ/ and /ɔ/ was to perceive them as /ʌ/. As a consequence, /ʌ/ showed a large increase in false identification. (3) The vowels /æ/ and /ʌ/ were also more ambiguous in destressed syllables. They were most frequently misperceived as /ɛ/ and /ʊ/, respectively. (4) In exception to this general pattern of increased error rates, the vowels /ɛ/ and /ʊ/ showed substantially fewer errors in destressed syllables. However, both vowels were popular false responses, and the apparent improvement was associated with a positive bias in each case. It is relevant that /ɛ/ and /ʊ/ are the most "central" vowels in two-formant space, in that they are intermediate in first-formant frequency, and therefore reduction toward schwa does not tend to produce formant combinations typical of other vowels. The tendency for listeners to select more "central" vowel responses suggests that they underestimated the tempo at which the excised syllables were spoken.

TABLE 3: Mean percent error in identification of destressed /p-p/ syllables.

Intended vowel	Condition: No-precursor	Point-vowel precursor	Sentence context
i	11.5	11.2	6.7
ɪ	0.5	1.8	0.7
ɛ	7.9	3.5	20.0
æ	24.5	44.1	2.0
ɑ	62.5 (43.0)	95.9 (92.4)	36.7 (12.7)
ɔ	49.5 (25.5)	50.6 (45.9)	31.3 (4.0)
ʌ	33.0	27.6	33.3
ʊ	19.0	18.2	23.3
u	4.5	4.7	1.3
Overall	23.8 (18.9)	28.6 (27.7)	17.3 (11.6)

Rather than enabling listeners to compensate for errors introduced by tempo uncertainty, the point-vowel precursors served only to increase the errors (see Table 3 and the confusion matrix in Table A-8). Listeners tended to hear vowels more centralized than those intended, and did so with even greater frequency than in the No-Precursor condition. The trend was so strong for /ɑ/ and /ɔ/ that confusions between them accounted for only 6 percent of errors on the two vowels themselves and only 3 percent of all errors on the Point-Vowel Precursor test. Relatively low error rates occurred on the two most "central" vowels, /ɛ/ and /ʊ/, as was found on the No-Precursor test.

It seems likely that the precursor syllables (spoken in citation form) established an expected tempo inappropriate for perception of the subsequent test syllables. Instead of calibrating listeners to the formant ranges of a talker's vowel space, the precursors calibrated listeners to the tempo of the talker's speech. If the test syllable had truly been spoken in isolation with a stress equal to that of the precursors, the prior adjustment to talker tempo would have been appropriate. This condition was met in the Point-Vowel Precursor test of Experiment II, where errors averaged only 15 percent. However, the comparable test in Experiment III juxtaposed syllables spoken with radically different rates and stresses, and the contrast produced a large increase in erroneous judgments. As in the No-Precursor condition, the pattern of errors reflected the contraction of acoustic vowel space found for rapid, destressed speech (cf. Lindblom, 1963).

In contrast to the results following precursors, error rates for individual vowels dropped when the destressed test syllables were heard in sentence context (see Table 3 and the confusion matrix in Table A-9). Error rates for /i/, /æ/, /ɑ/, /ɔ/, and /u/ were all lower in the Sentence Context condition than in the No-Precursor condition, where the syllables were heard in isolation. While errors on /ɛ/ and /ʊ/ were relatively infrequent in the excised syllables, they increased when heard in sentence context. In general, the pattern of changes was complementary to that observed for the excised syllables. The marked "centralization" of vowel responses disappeared when syllables were heard in sentence context.

These results suggest that a carrier sentence aids identification of vowel targets by allowing listeners to adjust to talker tempo, rather than by allowing them to compensate for talker variation. The observed changes in identification have little in common with those found after extended familiarization with a talker's speech (cf. Figure 2). When errors in the Sentence Context and No-Precursor conditions were compared, there were no vowels that showed "true improvement" in identification. The main effect of sentence context was to reverse a pattern of positive biases toward /ɛ/ and /ʊ/--and to a lesser extent /ɪ/ and /ʌ/--a pattern that has more to do with tempo uncertainty than with talker variation.

Luce analyses for the three experimental conditions corroborate the conclusions drawn from the less formal error analyses. Most pairwise confusions were greater for destressed syllables (No-Precursor condition) than for citation-form syllables (Mixed-Talker condition, Experiment II). In two cases, /ɑ-ɔ/ and /ɔ-ʌ/, the increases were large and significant. Thus, tempo uncertainty produced some genuine increases in vowel confusability. However, one significant decrease was also observed: the /ɛ-æ/ confusion, largest source of errors on citation-form syllables, was substantially smaller for rapid, destressed syllables. It is possible that rapid articulation produced tokens of /ɛ/ that would also have been produced with high probability in citation form--that is, rapid articulation may affect /ɛ/ more by reducing its acoustic variance than by shifting its typical formant composition. If this effect were large enough, the overall discriminability of /ɛ/ and /æ/ would increase, as observed.

Pairwise confusions for the Point-Vowel Precursor condition showed little systematic change relative to the No-Precursor condition. The only significant change was an increase in the confusability of /ɑ/ and /ʌ/. The /ɛ-æ/ confusion was more asymmetric than in the No-Precursor condition (/ɛ/ was never perceived as /æ/ following precursors), and the similarity showed a further, though nonsignificant decrease.

Pairwise confusions in the Sentence Context condition tended to be lower than in the No-Precursor condition, though only one of the decreases (/ɔ-ʌ/) was significant. Thus, sentence context reversed one of the two significant increases in confusability found for the excised syllables. The other vowel pair /ɑ-ɔ/ also showed a reversal, but the decrease was not significant.

While the observed changes in pairwise similarities were usually in the expected direction, they were also few in number. The predominant effect of misperceiving tempo was not a change in vowel similarities, but an error-producing shift in response biases. Joint Luce models for the citation-form syllables (Mixed Talker condition, Experiment II) and destressed syllables (No-Precursor condition) verify that the main impact of tempo uncertainty was on response biases. A same-parameters model ($\chi^2/df = 6.14$) was not improved by different similarity parameters ($\chi^2/df = 7.36$), but was substantially improved by different biases ($\chi^2/df = 3.86$). Joint Luce models comparing the destressed syllables in isolation (No-Precursor condition) with those in sentence context yield similar results: a same-parameters model ($\chi^2/df = 4.18$) was not improved by different similarities ($\chi^2/df = 6.58$), but was improved by different biases ($\chi^2/df = 2.27$). Again, these results for the Sentence Context condition contrast sharply with those for the Segregated Talker test (Experiment II), where the predominant effect was on pairwise similarities, not biases.

It is interesting to note that the error rate for syllable-medial vowels in sentence context (17.3 percent) was very close to that for medial vowels in citation-form syllables (17.0 percent); the difference was not significant [$t(32) = 0.16$]. This suggests that there is a very stable level of error for vowels in /p-p/ words when heard in a unit of articulation sufficient to specify tempo. The only additional assumption required is that a syllable spoken in isolation specifies its own tempo.

These results provide strong evidence that the perceptual system adjusts to the ongoing tempo of a talker's utterance. However, it remains an open question whether this adjustment involves transforming or calibrating a relational vowel space for individual talkers. No evidence for a talker-specific space of this kind was found in earlier experiments, nor was any found in the precursor condition of this experiment. In addition, the effect of sentence context on identification was very different from the effect of extended familiarization with individual vocal tracts. Thus, this experiment provides no evidence that sentence context aids vowel identification by allowing compensation for talker differences.

Little is currently known about how formant contours are transformed by variations in speaking rate and stress, or how listeners adjust to these changes. Lindblom (1963) has attempted to characterize the variation in vowel center formant frequencies as a function of speaking rate. Lindblom and Studdert-Kennedy (1967), in turn, have demonstrated that listeners are sensitive to these variations when identifying vowels in isolated, synthetic syllables. If two syllables reach the same formant frequency values at the syllable centers, but simulate different rates of articulation, listeners adopt different criteria for identification of the two medial vowels. These preliminary efforts suggest that the formant transitions, which are generally understood to carry consonantal information, must also aid in specifying the vowel. They apparently do so, at least in part, by limiting the range of possible talker tempos. The Sentence Context condition of this experiment suggests that factors beyond the syllable also shape the acoustic specification of vowels and are therefore important to accurate identification. A major function of a carrier sentence is to specify the tempo and stress of component syllables.[8]

SUMMARY AND CONCLUSIONS

These experiments lead to the following conclusions about the perception of vowels in natural speech:

---

[8] Gay's (1974) acoustic measurements suggest that the critical feature of destressed syllables in natural sentences is that they are destressed, not that they are rapidly spoken. Point vowels in rapidly spoken syllables did not show the reduction toward schwa that is found in destressed speech (Lindblom, 1963). It is not clear what implications this has for the perceptual studies of Lindblom and Studdert-Kennedy (1967) or the studies presented here. In both cases, tempo variation has provided a plausible basis for explanation. Further research is needed to determine whether perceived pace and syllable duration are secondary to perceived stress in determining the pattern of listeners' identifications.

1. Talker-dependent acoustic variation does not pose a major perceptual problem within a common dialect group. Listeners can identify a high proportion of vowels spoken in citation-form syllables by talkers with whom they have little or no previous experience. In Experiment I, listeners identified 87 percent of /h-d/ syllables spoken in random order by 30 talkers representing the full natural range of acoustic variation. In Experiment II, they identified 83 percent of /p-p/ syllables spoken by 15 talkers. Of the errors made in this Mixed Talker condition, no more than half can be attributed to talker-dependent sources of ambiguity. Correct identification in Segregated Talker tests averaged 90.5 percent for vowels in /p-p/ syllables (Experiment II). There was genuine improvement in the identification of specific vowels, but only a small portion of correct identification could be attributed to familiarization (the difference between 83 and 90.5 percent). Thus, experience with a voice plays a secondary role in specifying vowel identity. A single syllable contains substantial information about its medial vowel, whether a talker's voice is familiar or not.

2. Contrary to the speculations of Joos (1948), Lieberman et al. (1972), and Lieberman (1973), the point vowels do not play a major and privileged role as calibrators of a talker-specific vowel space. Experience with a talker's point vowels does not significantly reduce the overall ambiguity of vowels in a subsequent syllable. This result was found for all three types of test syllables studied: /h-d/, citation-form /p-p/, and destressed /p-p/. The pattern of changes following point-vowel precursors did not resemble the pattern resulting from extended experience with a talker's voice (Experiment II). Extended experience produced consistent reductions in pairwise similarities, while experience with a talker's point vowels mainly affected the pattern of response biases, with no consistent effects on vowel identifiability. Point vowels did produce a significant decrease in the confusability of /pɛp/ and /pæp/, but they were not unique in this respect: a significant reduction was also found when test syllables were preceded by central vowels (Experiment II) and when tempo uncertainty was introduced (Experiment III). In general, there was little evidence that sample subsets of a talker's vowels enable listeners to adjust to the talker's idiosyncratic "space" (defined by ranges of acoustic values or by sizes of vocal-tract cavities). This conclusion, like the first, does not support the proposal of Ladefoged and Broadbent (1957) and Ladefoged (1967) that vowel perception can be regarded as a problem in establishing an adaptation level (cf. Shankweiler, Strange, and Verbrugge, in press).

3. Listeners adjust their perceptual criteria for syllable-medial vowels according to the perceived rate of articulation. When destressed /p-p/ syllables were excised from sentence context and presented in isolation (Experiment III), there was a tendency to perceive them as if they had been spoken in citation form: the pattern of errors showed insufficient compensation for the acoustic effects of rapid articulation. When citation-form precursor strings preceded the excised syllables, the contrast of expected and actual tempos enhanced the original pattern of errors and increased the overall error rate. When the excised syllables were heard in their original temporal environments (the carrier sentences), the pattern of errors reversed and the overall error rate decreased. Carrier sentences apparently enabled listeners to adjust continuously to a talker's tempo and to compensate for the acoustic effects of vowel reduction. Information about a talker's ongoing tempo produced a qualitatively different pattern of improvement from that produced by long-term familiarization with citation-form syllables. This confirmed the results of Experiment II (where citation-form test words were heard in the context of prior citation-form syllables)

in the more natural situation of words in sentence context. In neither case was there evidence that listeners acquired a scaling function for adjusting a talker's speech to a normative dialectal space. In contrast to the conclusions of Ladefoged and Broadbent (1957), a naturally produced carrier sentence may aid vowel identification more by establishing the tempo of speech than by delimiting an individual vowel space.

How do listeners cope with talker-related acoustic variation? One possibility is that a single syllable (with consonants of known identity) carries sufficient information for normalization to take place. Fourcin (1968) and Rand (1971) both have demonstrated that listeners adjust their perceptual criteria for stop consonants to compensate for talker-dependent variation in the consonants' acoustic structure. If the consonants in a test syllable are known in advance, a single syllable could provide relatively unambiguous information about the talker's vocal tract. This information, in turn, could be used in disambiguating the vowel.

A second possibility is that a talker-normalization procedure is not necessary for human perception of vowels. Vowel identity may be specified by properties of the acoustic signal that are relatively invariant across talkers and that do not require a prior calibration process to be accurately detected. The results for destressed syllables suggest that the dynamic properties of speech are especially critical: vowel identification seems to be at least as sensitive to tempo variation as it is to variation in talkers' center formant frequencies. Adjustment to talkers may have more to do with tracking the dynamics of ongoing articulation than with normalization as traditionally defined.

REFERENCES

Abramson, A. S. and F. S. Cooper. (1959) Perception of American English vowels in terms of a reference system. Haskins Laboratories Quarterly Progress Report QPR-32, Appendix 1.

Fourcin, A. J. (1968) Speech source inference. IEEE Trans. Audio Electroacoust. AU-16, 65-67.

Gay, T. (1974) A cinefluorographic study of vowel production. J. Phonetics 2, 255-266.

Gerstman, L. H. (1968) Classification of self-normalized vowels. IEEE Trans. Audio Electroacoust. AU-16, 78-80.

Goodman, L. A. (1969) How to ransack social mobility tables and other kinds of cross-classification tables. Am. J. Sociol. 75, 1-40.

Goodman, L. A. (1970) The multivariate analysis of qualitative data: Interactions among multiple classifications. J. Am. Stat. Assoc. 65, 226-256.

Helson, H. (1948) Adaptation level as a basis for a quantitative theory of frames of reference. Psychol. Rev. 55, 297-313.

Joos, M. A. (1948) Acoustic phonetics. Language, Suppl. 24, 1-136.

Ladefoged, P. (1967) Three Areas of Experimental Phonetics. (New York: Oxford University Press).

Ladefoged, P. and D. E. Broadbent. (1957) Information conveyed by vowels. J. Acoust. Soc. Am. 29, 98-104.

Lieberman, P. (1973) On the evolution of language: A unified view. Cognition 2, 59-94.

Lieberman, P., E. S. Crelin, and D. H. Klatt. (1972) Phonetic ability and related anatomy of the newborn, adult human, Neanderthal man, and the chimpanzee. Am. Anthropol. 74, 287-307.

Lindblom, B. E. F. (1963) Spectrographic study of vowel reduction. J. Acoust. Soc. Am. 35, 1773-1781.

Lindblom, B. E. F. and M. Studdert-Kennedy. (1967) On the role of formant transitions in vowel recognition. J. Acoust. Soc. Am. 42, 830-843.

Lindblom, B. E. F. and J. Sundberg. (1969) A quantitative model of vowel production and the distinctive features of Swedish vowels. Quarterly Progress and Status Report (Speech Transmission Laboratory, Royal Institute of Technology, Stockholm, Sweden) STL-QPSR 1, 14-32.

Luce, R. D. (1959) Individual Choice Behavior: A Theoretical Analysis. (New York: Wiley).

Luce, R. D. (1963) Detection and recognition. In Handbook of Mathematical Psychology, ed. by R. D. Luce, R. R. Bush, and E. Galanter. (New York: Wiley), Vol. 1, pp. 103-189.

Peterson, G. E. (1961) Parameters of vowel quality. J. Speech Hearing Res. 4, 10-29.

Peterson, G. E. and H. L. Barney. (1952) Control methods used in a study of the vowels. J. Acoust. Soc. Am. 24, 175-184.

Peterson, G. E. and I. Lehiste. (1960) Duration of syllable nuclei in English. J. Acoust. Soc. Am. 32, 693-703.

Rand, T. C. (1971) Vocal tract size normalization in the perception of stop consonants. Haskins Laboratories Status Report on Speech Research SR-25/26, 141-146.

Shankweiler, D., W. Strange, and R. R. Verbrugge. (in press) Speech and the problem of perceptual constancy. In Perceiving, Acting, and Knowing: Toward an Ecological Psychology, ed. by R. Shaw and J. Bransford. (Hillsdale, N. J.: Lawrence Erlbaum Assoc.).

Shearme, J. N. and J. N. Holmes. (1962) An experimental study of the classification of sounds in continuous speech according to their distribution in the formant 1 - formant 2 plane. In Proceedings of the Fourth International Congress of Phonetic Sciences. (The Hague: Mouton), pp. 234-240.

Stevens, K. N. (1972) The quantal nature of speech: Evidence from articulatory-acoustic data. In Human Communication: A Unified View, ed. by E. E. David, Jr., and P. B. Denes. (New York: McGraw-Hill), pp. 51-66.

Stevens, K. N. and A. S. House. (1963) Perturbation of vowel articulations by consonantal context: An acoustical study. J. Speech Hearing Res. 6, 111-128.

Strange, W., R. R. Verbrugge, D. P. Shankweiler, and T. R. Edman. (in press) Consonant environment specifies vowel identity. J. Acoust. Soc. Am.

Tiffany, W. R. (1959) Nonrandom sources of variation in vowel quality. J. Speech Hearing Res. 2, 305-317.

Verbrugge, R., W. Strange, and D. Shankweiler. (1974) What information enables a listener to map a talker's vowel space? Haskins Laboratories Status Report on Speech Research SR-37/38, 199-208.

APPENDIX A: CONFUSION MATRICES

Tables report the frequency with which each intended vowel $\underline{x}$ was identified as response alternative $\underline{y}$. In addition, summary statistics for each condition are provided: the percent error for each intended vowel, the overall percent error for each repetition (rep.) of the test series, the overall percent error pooling both repetitions, the total number of trials for the two repetitions, the mean number of trials on which listeners made an error ($\bar{x}$), the standard deviation of this mean (s), and the number of listeners (N).

TABLE A-1: /h-d/ syllables: No-Precursor condition.[a]

Intended vowel	Response																Percent error
	i	ɪ	ɛ	æ	ɑ	ɔ	ʌ	ʊ	u	ɝ	eɪ	oʊ	aɪ	aʊ	ɔɪ	None	
i	202		1												1		1.0
ɪ	1	163	39										1				20.1
ɛ		12	165	24			1			2							19.1
æ			4	179	12	1	2			1	1		1	1	1	1	12.3
ɑ				5	105	80	4				1		4	2	2	1	48.5
ɔ					18	167	10	1				1		4	3		18.1
ʌ			1	1	24	2	174						2				14.7
ʊ					7		14	174	1		1	4			3		14.7
u					1			13	187			3					8.3
ɝ										204							0.0
eɪ			5								199						2.4
oʊ					1			1	16	1	3	178		1	1	2	12.7
aɪ		2			1						1		200				2.0
aʊ				1	1	3						26		171	1	1	16.2
ɔɪ					1					7					196		3.9

[a]Overall percent error: 12.94 (pooled), 14.97 (rep. 1), 10.92 (rep. 2); 180 trials, $\bar{x}$ = 23.29, s = 8.56, N = 17.

TABLE A-2: /h-d/ syllables: Point-Vowel Precursor condition.[a]

Intended vowel	Response																Percent error
	i	ɪ	ɛ	æ	ɑ	ɔ	ʌ	ʊ	u	ɝ	eɪ	oʊ	aɪ	aʊ	ɔɪ	None	
i	240																0.0
ɪ		169	68								1	1	1				29.6
ɛ		3	218	17						1		1					9.2
æ			11	217	5	5	2										9.6
ɑ				4	136	93	6				1						43.3
ɔ		1		1	57	137	22				3	4		5	1	9	42.9
ʌ					7	2	231										3.8
ʊ					1	1	36	196	2		1	1		1	1		18.3
u								1	236			3					1.7
ɝ										240							0.0
eɪ		1	2			1					235					1	2.1
oʊ						4		1	4			229			2		4.6
aɪ													240				0.0
aʊ				2	1	26					1	13		197			17.9
ɔɪ															240		0.0

[a]Overall percent error: 12.19 (pooled), 13.17 (rep. 1), 11.22 (rep. 2); 180 trials, $\bar{x}$ = 21.95, s = 5.79, N = 20.

TABLE A-3: Citation-form /p-p/ syllables: Mixed Talker condition.[a]

Intended vowel	Response i	ɪ	ɛ	æ	ɑ	ɔ	ʌ	ʊ	u	None	Percent error
i	188		1						1		1.1
ɪ		187	1			2					1.6
ɛ			139	47	3			1			26.8
æ			33	154		2				1	18.9
ɑ					152	19	17	2			20.0
ɔ				1	46	138	1	4			27.4
ʌ					18	5	161	6			15.3
ʊ		8			2		47	116	16	1	38.9
u							2	3	185		2.6

[a]Overall percent error: 16.96 (pooled), 18.48 (rep. 1), 15.44 (rep. 2); 90 trials, $\bar{x}$ = 15.26, s = 4.53, N = 19.

TABLE A-4: Citation-form /p-p/ syllables: Segregated Talker condition.[a]

Intended vowel	Response i	ɪ	ɛ	æ	ɑ	ɔ	ʌ	ʊ	u	None	Percent error
i	329	1									0.3
ɪ	3	318	4				2	2		1	3.6
ɛ	1		290	20	4	7	5			3	12.1
æ			5	324		1					1.8
ɑ				7	255	62	4	2			22.7
ɔ					55	269	2	4			18.5
ʌ					11	9	305	4		1	7.6
ʊ						29	19	272	10		17.6
u							1	2	327		0.9

[a]Overall percent error: 9.46 (pooled), 10.57 (rep. 1), 8.35 (rep. 2); 90 trials, $\bar{x}$ = 8.52, s = 4.77, N = 33.

TABLE A-5: Citation-form /p-p/ syllables: Point-Vowel Precursor condition.[a]

Intended vowel	Response i	ɪ	ɛ	æ	ɑ	ɔ	ʌ	ʊ	u	None	Percent error
i	145		5								3.3
ɪ		146	3							1	2.7
ɛ		1	143	4		1	1				4.7
æ			30	119			1				20.7
ɑ				1	85	25	36	3			43.3
ɔ				1	9	122	14	4			18.7
ʌ					3	7	136	4			9.3
ʊ						2	31	110	7		26.7
u								11	139		7.3

[a]Overall percent error: 15.19 (pooled), 17.48 (rep. 1), 12.89 (rep. 2);
90 trials, $\bar{x}$ = 13.67, s = 5.26, N = 15.

TABLE A-6: Citation-form /p-p/ syllables: Central-Vowel Precursor condition.[a]

Intended vowel	Response i	ɪ	ɛ	æ	ɑ	ɔ	ʌ	ʊ	u	None	Percent error
i	116	3								1	3.3
ɪ	1	118								1	1.7
ɛ			107	12						1	10.8
æ			22	98							18.3
ɑ					85	20	12	3			29.2
ɔ					13	104	1	1		1	13.3
ʌ					10	8	93	9			22.5
ʊ						6	24	85	5		29.2
u								7	113		5.8

[a]Overall percent error: 14.91 (pooled), 15.00 (rep. 1), 14.81 (rep. 2);
90 trials, $\bar{x}$ = 13.42, s = 3.78, N = 12.

TABLE A-7: Destressed /p-p/ syllables: No-Precursor condition.[a]

Intended vowel	Response i	ɪ	ɛ	æ	ɑ	ɔ	ʌ	ʊ	u	None	Percent error
i	177	16	6						1		11.5
ɪ		199							1		0.5
ɛ[b]		2	164	7		1	2		2		7.9
æ			48	151		1					24.5
ɑ					75	39	76	10			62.5
ɔ				2	48	101	43	6			49.5
ʌ		8	5		1	15	134	35	1	1	33.0
ʊ			1		1	2	23	162	12		19.0
u							2	7	191		4.5

[a]Overall percent error: 23.84 (pooled), 25.19 (rep. 1), 22.49 (rep. 2); 90 trials, $\bar{x}$ = 22.00, N = 9; 88 trials, $\bar{x}$ = 20.55, N = 11; pooled scores: $\bar{x}$ = 21.20, s = 4.98, N = 20.

[b]Two trials lost for 11 subjects.

TABLE A-8: Destressed /p-p/ syllables: Point-vowel Precursor condition.[a]

Intended vowel	Response i	ɪ	ɛ	æ	ɑ	ɔ	ʌ	ʊ	u	None	Percent error
i	151	6	1					2	10		11.2
ɪ		167						3			1.8
ɛ		2	164				3	1			3.5
æ			74	95		1					44.1
ɑ		1	2	1	7	6	151	2			95.9
ɔ					8	84	69	9			50.6
ʌ		1	3	1	1	13	123	25	3		27.6
ʊ					4	1	11	139	15		18.2
u					1		1	6	162		4.7

[a]Overall percent error: 28.63 (pooled), 28.89 (rep. 1), 28.37 (rep. 2); 90 trials, $\bar{x}$ = 25.76, s = 4.70, N = 17.

TABLE A-9: Destressed /p-ɒ/ syllables: Sentence Context condition.[a]

Intended vowel	Response										Percent error
	i	ɪ	ɛ	æ	ɑ	ɔ	ʌ	ʊ	u	None	
i	140	10									6.7
ɪ		149							1		0.7
ɛ			120	29		1					20.0
æ			2	147				1			2.0
ɑ				2	95	36	15	1	1		36.7
ɔ					41	103	3	3			31.3
ʌ				1	8	17	100	24			33.3
ʊ					1	4	20	115	10		23.3
u					1			1	148		1.3

[a]Overall percent error: 17.26 (pooled), 18.22 (rep. 1), 16.30 (rep. 2); 90 trials, $\bar{x}$ = 15.53, s = 5.08, N = 15.

Identification of Dichotic Fusions*

Bruno H. Repp+

ABSTRACT

Seven synthetic syllables from a "place continuum" (/bæ - dæ - gæ/) were presented in all dichotic combinations for identification. These syllables fused completely, so that dichotic pairs were perceived as single stimuli. The response pattern could not be easily explained by an "auditory averaging" hypothesis. Rather, stimuli that were good instances of a category seemed to "dominate" stimuli that were closer to a category boundary. To account for this finding, a three-stage pattern recognition ("prototype") model is proposed according to which the information from the two ears is integrated after auditory but before phonetic-categorical processing, at a "multicategorical" stage. Electronically mixed stimuli led to a similar response pattern, suggesting that competing transitional cues remain intact up to the multicategorical stage. It is demonstrated that these fusions cannot be reliably discriminated from binaural stimuli, and that selective attention to one ear has little effect. For the purpose of assessing ear advantages, dichotic fusions offer methodological advantages over other dichotic stimuli. The problem of determining the "true" ear advantage is discussed.

INTRODUCTION

In recent years, dichotic listening has received much attention, both as a research tool for the investigation of the processes involved in speech perception and as a diagnostic technique for assessing hemispheric dominance for

---

*A substantially revised version of this paper is to be published in the Journal of the Acoustical Society of America. Authors who wish to refer to this research are urged to consult the revised version.

+Also University of Connecticut Health Center, Farmington.

Acknowledgment: This research was conducted at Haskins Laboratories and would not have been possible without the extraordinary hospitality of this institution and its director, Alvin Liberman. I thank him, Michael Studdert-Kennedy, James Cutting, Terry Halwes, Gary Kuhn, and David Paul for comments and discussions related to this paper. The author was supported by NIH Grant T22 DE00202 to the University of Connecticut Health Center.

[HASKINS LABORATORIES: Status Report on Speech Research SR-45/46 (1976)]

speech.[1] Both aspects are addressed by this paper, which, on the basis of a detailed analysis of the dichotic interaction between the voiced stop consonants, makes recommendations for a possible methodological refinement of dichotic testing.

Dichotic tests composed of synthetic stop-consonant-vowel syllables have become widely accepted as the most precise instruments currently available for assessing ear advantages in speech perception (Shankweiler and Studdert-Kennedy, 1967a, 1975). The control of stimulus characteristics and channel synchronization made possible by modern speech synthesizers and specialized computer systems, together with the balanced stimulus set of the six stop consonants, gives these tests a distinct advantage over other materials and procedures. Nevertheless, some problems remain. One is the kind and number of responses to be required from the listeners: two responses (with or without restrictions on their order) or one response (with or without selective-attention instructions)? Variants of both response modes have been used at one time or another, but two-response paradigms have dominated the scene. However, because of the occurrence of confusions, intrusions, and guessing, and the lack of a good theory taking these phenomena into account, the two responses cannot be unequivocally assigned to the stimuli that evoked them, so that errors and correct responses are not clearly separated in scoring the results (cf. Repp, 1975a, 1976). Selective-attention instructions offer no remedy, since selective attention is very difficult with precisely aligned dichotic syllables, and intrusions from the unattended channel are common (Halwes, 1969; Haggard, 1975; Repp, 1975a).

Another problem has been the derivation of an idex for the ear advantage. Simple percentage differences have the disadvantage that they depend on the overall performance level and therefore do not adequately represent the degree of an ear advantage but merely measure its direction. The proposal of Kuhn (1973) to use the ϕ coefficient as a measure of the ear advantage has been an important step forward. However, Kuhn's index is designed for two-response paradigms (or single-response paradigms with selective-attention instructions) and therefore does not solve the problem of unraveling correct responses and errors.

Halwes (1969) and Studdert-Kennedy and Shankweiler (1970) have pointed out the low information content of the second of two responses. This observation suggests that it may be more appropriate to ask for a single response only. In fact, it seems that listeners often perceive only a single syllable when a dichotic pair is presented. This tendency is more pronounced with syliables contrasting in only a single distinctive feature[2] (voicing, for example,

---

[1] See, for example, Brain and Language, 1974, Vol. 1, No. 4 and 1975, Vol. 2, No. 2.

[2] A comment on terminology is in order here. Many authors refer to "shared features" rather than "feature contrasts," for example, /ba/ and /pa/ "share place" (Studdert-Kennedy and Shankweiler, 1970; Pisoni and McNabb, 1974). This terminology is awkward, for several reasons: (1) Any characterization in terms of shared features is indeterminate unless all shared features are enumerated (which includes many irrelevant features), whereas mentioning the contrasting features is informative even without precise knowledge of the complete stimulus set. (2) Features are dimensions and therefore are always shared, precisely

/ba+pa/; or place, for example, /ba+da/)[3] than with syllables contrasting in both features (for example, /ba+ta/): in a "same-different" judgment task, the former receive more incorrect "same" responses than the latter. Moreover, within the single-feature contrasts, place contrasts are much harder to discriminate from identical (binaural) syllables than voicing contrasts (Halwes, 1969; Blumstein and Cooper, 1972; Repp, 1976). In other words, precisely aligned simultaneous dichotic syllables that differ only in the direction of their initial formant transitions strongly tend to fuse and sound like a single syllable originating in the middle of the head (if their intensities are equal).

Cutting (1972, 1976) has proposed a classification of dichotic fusions that includes "psychoacoustic fusions": when /ba+ga/ is presented, /da/ is often heard. We will follow Cutting and use the term "psychoacoustic fusion" only for this specific phenomenon. However, it should be clear that fusion in the more general sense--hearing only a single stimulus when two are presented--occurs independently of the nature of the phonetic percept.[4] Thus, /ba+ga/ sounds just as fused when /ba/ or /ga/ is heard as when /da/ is heard, and /ba+da/ fuses just as well, although it will never give rise to a "new" response.

These considerations suggest that it is useless to require a listener to give two responses when a dichotic place contrast is presented. A single response will contain virtually all the information available to the listener. (However, it may be usefully supplemented by a measure of response uncertainty, such as confidence ratings, reaction times, or response distributions.) The principal question is then: How is the information from the two ears combined into a single percept? Cutting (1972, 1976) has suggested that psychoacoustic fusion is a relatively low-level auditory averaging phenomenon. Any such explanation should apply to all dichotic place contrasts. The present experiments attempt to investigate this question further by examining the identification of dichotic fusions in some detail.

From a methodological standpoint, it is important to determine whether dichotic fusions lead to the right-ear advantage (REA) commonly found in dichotic listening. Several studies have indicated that place contrasts show a somewhat

---

speaking. It is their values that may differ, and this seems to be somewhat better captured in the term "feature contrast" (that is, a contrast with respect to a feature) than in "shared feature." (3) Most importantly, feature sharing has often been interpreted as a factor facilitating dichotic perception. However, there is no known factor in dichotic listening that facilitates perception relative to monaural or binaural presentation; rather, performance is impaired by competition as a consequence of feature contrasts. Therefore, the latter term will be used here exclusively.

[3]The notation i+j will be used to indicate a dichotic stimulus pair regardless of channel/ear assignment of the component stimuli, while i-j and j-i will designate the two specific channel assignments (i and j stand for stimulus numbers; see Table 1).

[4]Conversely, it may also be argued that, within the set of the six stop consonants at least, there is characteristically only one perceptual result, regardless of whether phenomenological fusion occurs.

smaller REA than other feature contrasts (Shankweiler and Studdert-Kennedy, 1967a; 1967b; Studdert-Kennedy and Shankweiler, 1970). Since the place contrasts in these studies may not have been perfectly fused, the difference may in fact be larger. This is interesting with regard to the question at which level(s) in processing the REA arises. If it were the case that dichotic place contrasts fuse at a very early stage in processing and then are transmitted in this form to each hemisphere, there should be no REA, since the REA is usually attributed to transcallosal transmission loss of left-ear information, assuming functional independence of the dichotic inputs prior to their convergence upon the dominant hemisphere (Studdert-Kennedy, 1975). On the other hand, fusion may either occur at a higher level (after central convergence) or be an entirely autonomous phenomenon mediated by an independent low-level cross-correlational mechanism, so that fused syllables are processed in basically the same way as less completely fused syllables; in this case, there should be no difference in REAs between the two.

EXPERIMENT I

The first experiment examined the identification of fused dichotic stimuli from a "place continuum" (Pisoni, 1971) obtained by systematically varying the starting frequencies of the initial formant transitions. The principal questions were whether identification responses could be predicted by a simple auditory averaging model, whether a significant REA of "normal" magnitude exists, and whether psychoacoustic fusions are as common as suggested by Cutting (1972).[5] The effects of variations in the acoustic properties and relationships of the fused stimuli were of prime concern with respect to all three questions.

Method

Subjects. Thirteen paid volunteers participated, seven males and six females, all right-handed, unaware of any hearing trouble, and relatively inexperienced listeners. The data of two additional subjects were eliminated because they were too noisy.

Stimuli. The stimuli were seven syllables ranging perceptually from /bæ/ to /dæ/ to /gæ/. They were produced on the Haskins Laboratories parallel resonance synthesizer. All syllables were of 280-msec duration, had a constant fundamental frequency (114 Hz), a voice onset time of -15 msec (that is, prevoicing), 45-msec linear transitions, and no bursts but an abrupt onset of energy following the prevoicing. The syllables differed only in the onset frequencies of the second-formant (F_2) and third-formant (F_3) transitions, which are shown in Table 1.

Dichotic pairs were constructed using the pulse code modulation (PCM) system at Haskins Laboratories. The stimulus alignment precision of this computerized procedure is ±0.125 msec. All possible combinations of the seven stimuli were recorded. In order to obtain stable identification scores for the seven syllables in isolation (that is, binaurally), pairs of identical syllables were replicated six times, so that there were 84 stimuli altogether: 42 identical

---

[5] The recent paper of Cutting (1976) was not available at the time of the experiment.

TABLE 1: Starting frequencies (in Hz) of second-formant (F_2) and third-formant (F_3) transitions of the seven stimuli.

Stimulus Number	F_2	F_3
1	1312	2348
2	1456	2694
3	1620	3026
4	1772	3026
5	1920	2694
6	2078	2348
7	2234	2018
Steady-state /æ/	1620	2862

(binaural) pairs and 42 nonidentical (dichotic) pairs. Five different random sequences of the 84 stimuli were recorded. The interstimulus interval was 3 sec.

Procedure. The subjects were tested individually or in small groups in a single session lasting approximately 90 minutes. Playback was from an Ampex AG-500 tape recorder through an amplifier to Grason-Stadler TDH-39 earphones. Playback intensity was adjusted and monitored on a Hewlett-Packard voltmeter, and special care was taken to equalize the intensities of the two channels at about 85 dB SPL (peak deflections).

Each subject listened twice to the five blocks of 84 stimuli. The channels were reversed electronically after the first five blocks. The instructions were to write down one response for each syllable heard: B, D, or G, whatever the syllable sounded most like.

The subjects were generally not informed until after the experiment that different inputs were presented to the two ears in half of the stimuli. (There were some exceptions, because some subjects had previously participated in related experiments with dichotic fusions.) Most subjects agreed when questioned that they heard only single syllables and showed surprise when told about their actual nature. This, together with the experimenter's impression, was considered sufficient evidence for the adequate fusion of the stimuli. (Formal tests were conducted later in Experiment III with different subjects.)

Results and Discussion

The response pattern. The pooled results of the 13 subjects are shown in Figure 1. The numbers in the graphs represent identical (binaural) pairs, and the dashed lines connecting them trace the categorical identification functions for the seven stimuli. It can be seen that stimuli 1 and 2 were generally identified as B; 3 and 4, as D; and 6 and 7, as G. Stimulus 5 was the only truly ambiguous syllable, with somewhat more D than G responses. (The stimulus numbers refer to Table 1.) Some subjects produced noisy data, which is reflected in the averages; for example, G responses to stimuli 6 and 7 reached only 85-86 percent.

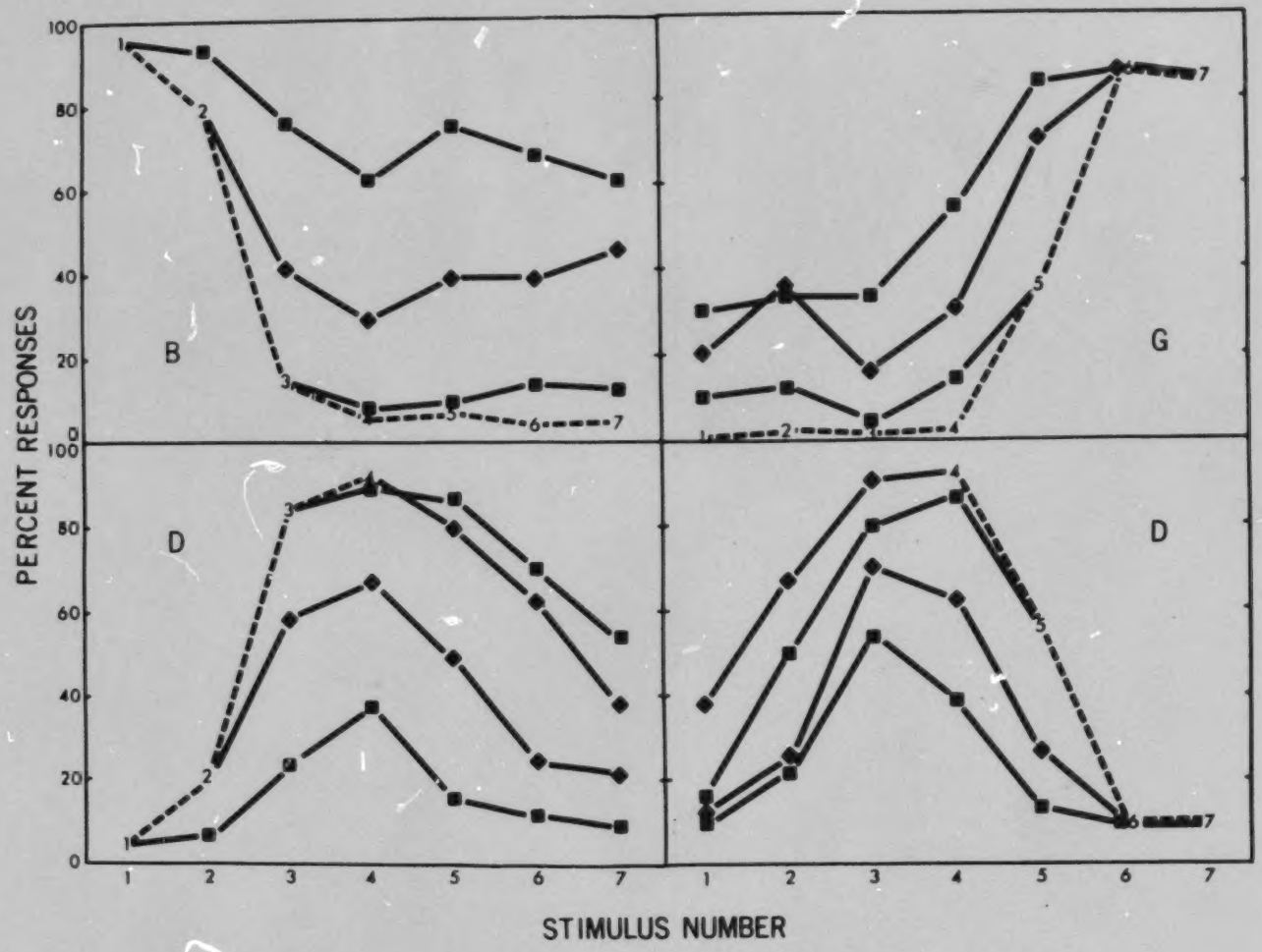


FIGURE 1

Figure 1: Percentages of B responses (upper left-hand panel), G responses (upper right-hand panel), and D responses (lower panels) to binaural (numbers) and dichotic (filled symbols) pairs. Each function connects the combinations of a constant stimulus (represented by the number at one of the endpoints of the function) with the stimuli along the abscissa. The dashed lines connect the binaural pairs and trace the single-stimulus labeling functions.

Consider now the other symbols in Figure 1 that represent the dichotic combinations of different stimuli. Each function connects the pairs formed by one particular stimulus (denoted by the number at one end of the function) and the stimuli along the abscissa. The pattern may be described as follows:

(1) When a particular stimulus was paired with other stimuli, the percentage of responses in the relevant category tended to decrease as the competing stimuli were further and further removed on the continuum. This was especially clear for D responses, while the functions for B and G responses became flat and even nonmonotonic when /bæ/ and /gæ/ stimuli were paired with stimuli more than two or three steps removed on the continuum. Note that B responses were at a minimum in pairs with stimulus 4, while G responses tended to be at a minimum in pairs with stimulus 3.

(2) The percentages of responses in the three categories generally remained in proportion to the binaural identification results for the component syllables of a dichotic pair; for example, the B-function for stimulus 2 (upper left-hand panel in Figure 1) lies uniformly lower than that for stimulus 1, and that for stimulus 3 is even lower. More interesting, however, is the fact that a similar difference exists between the G-functions for stimuli 6 and 7 (upper right-hand panel in Figure 1), although these two stimuli showed identical binaural identification scores. In addition, there is one crossover of functions: the D-function for stimulus 3 lies above that for stimulus 4 in pairs with stimuli 5, 6, and 7 (lower left-hand panel in Figure 1).

(3) There was a tendency for /bæ/ stimuli (especially 1) to dominate /gæ/ stimuli (6 and 7). A change in acoustic structure at the /bæ/ end of the continuum had a greater effect than an equivalent change at the /gæ/ end, as indicated by the wider spacing of the B functions (cf. upper panels in Figure 1).

(4) Psychoacoustic fusions were clearly present but rather infrequent, especially in pairs containing stimulus 1. The numerical results for the four relevant stimulus pairs are shown in Table 2.

TABLE 2: Response percentages for four dichotic /bæ + gæ/ pairs (13 subjects).

Stimulus pair	Responses		
	B	D	G
2 + 6	38.5	25.4	36.1
2 + 7	45.0	21.5	33.5
1 + 6	67.3	12.7	20.0
1 + 7	60.8	9.6	29.6

Psychoacoustic fusions: Three averaging hypotheses. The pattern of results just described (particularly under paragraphs 2 and 4) definitely rules out a "phonetic averaging" (attention-switching or rivalry) hypothesis. If, for example, the two stimuli competed for a single phonetic processor, so that one syllable gained access to the processor in a certain percentage of the trials while the other syllable was lost, the distribution of identification responses

for a dichotic pair would be a weighted average of the response distributions for the two component stimuli in isolation. The same would be true if both syllables were categorized independently in separate processors, and an attentional mechanism with limited capacity selected one or the other outcome on a probabilistic basis. Instead, the existence of psychoacoustic fusions and of effects of acoustic within-category differences is evidence that the dichotic information interacts prior to the completion of phonetic processing.

A second hypothesis may be termed "articulatory averaging" (Cutting, 1976). It is similar to the phonetic averaging hypothesis, except that it allows for psychoacoustic fusions by perceptual-articulatory interpolation at the feature level. However, it excludes any interaction between the acoustic properties of the stimuli and therefore is clearly disconfirmed both by the present data and by Cutting's own.

On the other hand, the data are superficially in accord with an "auditory averaging" hypothesis, which assumes that the formant transitions of the two competing stimuli (or rather, their equivalent auditory codes in the brain) fuse to yield new, intermediate transitions, and the resulting new information is then phonetically interpreted. This hypothesis has also been considered by Cutting (1976), who independently investigated the effect of acoustic stimulus variations on the frequency of psychoacoustic fusions. However, one prediction would then be that /bæ + gæ/ stimulus pairs, such as 1 + 7 and 2 + 6, which have about the same "average," should yield the same percentage of D responses. Instead, the acoustically more similar pair, 2 + 6, led to more psychoacoustic fusions than the acoustically more dissimilar pair, 1 + 7 (cf. Tables 2 and 3, upper left-hand quadrant), which parallels the results of Cutting (1976). Therefore, Cutting's conclusion that simple averaging of formant transitions is an insufficient explanation also applies to the present data.[6]

Another problem with the auditory averaging model is its deterministic nature. There is no /bæ + gæ/ stimulus pair for which only D responses are obtained. In fact, the frequency of psychoacoustic fusions in the present experiment was surprisingly low. Nine of the thirteen subjects showed negligible frequencies (less than 7 percent, after a correction for expected confusions). One reason for this may have been the presence of F_3 transitions, which were rising for /bæ/ and /gæ/ stimuli but falling in /dæ/ stimuli. In /bæ + gæ/ pairs, the "average" F_2 transition may have been in conflict with the "average" F_3 transition, so that the responses tended to shift among all three alternatives. The classical studies of Harris, Hoffman, Liberman, Delattre, and Cooper (1958) and Hoffman (1958) have shown (incidentally, also in the context /-æ/) that F_3 transitions have a strong influence on the tendency to give D responses, with F_2 transitions held constant: rising transitions decrease and falling transitions increase D responses. Cutting (1976) used two-formant syllables and obtained higher percentages of psychoacoustic fusions than the present study;

---

[6] Of course, the assumption of a linear (unweighted) auditory averaging process is naive and probably wrong. However, the conclusion that acoustic similarity plays a role seems nevertheless justified. The present results differ from those of Cutting (1976) with respect to the relative weight of low-frequency and high-frequency transitions. Here, low-frequency changes had a greater effect, while Cutting's data (for /ba + ga/) show precisely the opposite.

however, he also encouraged D responses by presenting only /ba + ga/ pairs to uninformed subjects who were given three response alternatives.

In order to check further on the role of F_3 transitions, a new stimulus tape was prepared that contained all dichotic and binaural pairs of seven syllables identical with those of Experiment I, except that they had no third formant. BHR, who had also participated in five sessions of Experiment I, listened to 30 random blocks of 49 stimulus pairs each, in three sessions. The results closely resembled his results with three-formant syllables, except for two of the four /bæ + gæ/ combinations. These results are shown in Table 3 (upper portion). The pooled response distribution for the four two-formant /bæ + gæ/ pairs differed significantly from that for the corresponding three-formant pairs ($\chi^2_{(2)} = 7.6$, $p < .05$) but, clearly, the difference was due only to 2 + 6 and 2 + 7, which showed greatly increased frequencies of D responses. (Note that BHR generally gave an unusually high percentage of psychoacoustic fusion responses.)

TABLE 3: Response percentages for four dichotic /bæ + gæ/ pairs: comparison of three-formant and two-formant syllables in dichotic and mixed presentation (data for a single practiced subject, BHR, based on 3-5 sessions per condition).

Stimulus pair	Responses					
	Three formants			Two formants		
Dichotic	B	D	G	B	D	G
2 + 6	28.0	47.0	25.0	26.7	70.0	3.3
2 + 7	36.0	38.0	26.0	23.3	53.4	23.3
1 + 6	71.0	26.0	3.0	73.3	23.4	3.3
1 + 7	63.0	30.0	7.0	71.7	25.0	3.3
Mixed						
2 + 6	55.0	25.0	20.0	36.7	58.3	5.0
2 + 7	96.3	2.5	1.2	86.7	11.7	1.7
1 + 6	46.3	38.8	15.0	30.0	56.7	13.3
1 + 7	77.5	10.0	12.5	80.0	6.7	13.3

It may be concluded that, in two pairs at least, the conflict between F_2 and F_3 transitions probably played a role. However, even in the absence of a third formant, psychoacoustic fusions were far from the 100 percent predicted by a simple auditory averaging hypothesis. If this hypothesis is to be maintained, considerable random variability in the weighting function of the averaging process must be assumed. This assumption will be tested in Experiment II.

Ear dominance and stimulus dominance. In order to correct for perceptual confusions between the stimulus categories (especially those provided by an ambiguous stimulus), left-ear and right-ear scores were derived for each stimulus pair. This was done by weighting each response by the relative frequencies of this particular response category for the two component stimuli in isolation and by subsequent summation of these weights for each ear. Expressed formally,

the right-ear score for a given dichotic pair i-j (with i in the right ear and j in the left ear) was computed as

$$T_{RE(i)} = \sum_{k=1}^{3} f(R_k|i\text{-}j)\,\frac{f(R_k|i)}{f(R_k|i) + f(R_k|j)} \qquad (1)$$

where $f(R_k|i\text{-}j)$ is the frequency of response category R_k for the dichotic pair, $f(R_k|i)$ and $f(R_k|j)$ are the frequencies of response R_k to i and j, respectively, when presented in isolation, and the summation is over the three response categories. For the left ear, $T_{LE(j)} = N - T_{RE(i)}$, where N is the total number of responses to this stimulus pair. The weight (the fraction) in Eq. (1) was set equal to 0.5 whenever the combined responses to i and j in a particular category constituted less than 10 percent. The resulting scores are free from overt variations in performance level, since the scores for the two ears always sum up to N, that is, there are no errors by definition. Because of the weighting procedure, individual variations in accuracy (which do exist) play only a negligible role as long as the "noise" does not exceed a certain level.

The two scores for a given dichotic pair, $T_{RE(i)}$ and $T_{LE(j)}$, have counterparts in the two scores for the other channel assignment of the same stimulus combination, $T_{RE(j)}$ and $T_{LE(i)}$. These four scores were arranged in two different two-way contingency tables, and two ø coefficients were calculated: the stimulus dominance index

$$\phi_D = (T_{RE(i)} - T_{RE(j)})/(T_{RE}T_{LE})^{1/2} \quad \text{with } T_{RE} = T_{RE(i)} + T_{RE(j)} \quad \text{and } T_{LE} = T_{LE(i)} + T_{LE(j)} \qquad (2)$$

which indicates the degree to which stimulus i "dominates" stimulus j; and the ear dominance (or ear advantage) index

$$\phi_E = (T_{RE(i)} - T_{LE(i)})/(T_{(i)}T_{(j)})^{1/2} \quad \text{with } T_{(i)} = T_{RE(i)} + T_{LE(i)} \quad \text{and } T_{(j)} = T_{RE(j)} + T_{LE(j)} \qquad (3)$$

which describes the relative dominance of the right ear over the left ear. Overall indices were obtained by calculating ø coefficients from summed response frequencies, with separate summations for i-j and j-i pairs (arbitrarily assuming that i<j on the stimulus continuum).[7] The significance of these indices was tested by $\chi^2_{(1)} = N\phi^2$ (cf. Kuhn, 1973).

[7] The denominator in the formula for the ø coefficient is the geometric mean of the two unequal marginal sums in the contingency table (the other two marginals being equal to N/2). Unless the difference between these marginals is very large, their geometric mean is similar to their arithmetic mean, which equals N/2. ϕ_D [Eq. (2)] is therefore usually well approximated by $2(T_{RE(i)} - T_{RE(j)})/N$, and ϕ_E [Eq. (3)] is usually almost identical to $2(T_{RE(i)} - T_{LR(i)})/N$, except in cases of extreme stimulus dominance. If the entries in the contingency table are expressed as percentages (that is, divided by N/2), ϕ_E and ϕ_D can be estimated at glance. This relationship also justifies the calculation of an overall index from summed response frequencies, which usually deviates only very slightly from the average of the coefficients for individual stimulus pairs.

The crucial question was whether the REA obtained from Eq. (3) would be comparable to the REA found in a two-response paradigm with a larger stimulus ensemble. The results are shown in the left third of Table 4. The 13 subjects exhibited a significant average REA, with six significant individual REAs but only one significant left-ear advantage. These results were compared with those of a recent study that used the complete set of six stop consonants and reported the distribution of Kuhn's (1973) ø coefficient for 22 subjects (Shankweiler and Studdert-Kennedy, 1975). The two distributions were virtually identical (Mann-Whitney test: $z = 0.03$). To the degree that the two ear-advantage indices are indeed equivalent, and within the limits imposed by the small sample sizes, this comparison indicates that dichotic fusions show just the same degree of an average REA as less completely fused syllables (which make up the majority of the combinations of all six stop consonants), so that phenomenological fusion is probably unrelated to the degree of REA obtained. The smaller REAs reported for place contrasts in the past were most likely artifacts of the two-response requirement and of the ear-advantage indices used.

TABLE 4: Dichotic ear dominance indices, and dichotic and mixed stimulus dominance indices for individual subjects.

	Dichotic ear dom.			Dichotic stim. dom.			Mixed stim. dom.		
Subject No.	ϕ_E	$\chi^2(1)$	p	ϕ_D	$\chi^2(1)$	p	ϕ_D	$\chi^2(1)$	p
1	0.13	7.1	<.01	0.10	4.4	<.05	0.04	0.7	n.s.
2	0.01	0.0	n.s.[a]	0.26	29.3	<.001	0.07	1.8	n.s.
3	-0.04	0.7	n.s.	-0.37	57.7	<.001	-0.48	98.1	<.001
4	0.06	1.6	n.s.	0.02	0.2	n.s.	(excluded)		
5	0.10	4.5	<.05	0.19	15.6	<.001	(excluded)		
6	-0.06	1.3	n.s.	0.19	15.2	<.001	-0.01	0.0	n.s.
7	0.14	7.8	<.01	0.14	8.4	<.01	0.22	20.8	<.001
8	0.10	4.1	<.05	0.13	7.5	<.01	0.30	37.0	<.001
9	-0.03	0.3	n.s.	-0.15	10.0	<.01	0.35	51.4	<.001
10	0.18	13.2	<.001	0.18	14.3	<.001	(no data)		
11	0.22	20.9	<.001	0.02	0.2	n.s.	(no data)		
12	0.06	1.5	n.s.	0.38	60.2	<.001	(no data)		
13	-0.12	6.5	<.02	0.03	0.5	n.s.	(no data)		
Total	0.06	17.9	<.001	0.09	41.7	<.001	0.07	14.0	<.001
BHR (3-F)[b]	0.11	25.2	<.001	0.05	4.9	<.05	(not calculated)		
BHR (2-F)[c]	0.03	1.0	n.s.	0.11	15.1	<.001	(not calculated)		

[a] $p > .05$.

[b] Three-formant syllables, calculated from the totals over five sessions.

[c] Two-formant syllables, calculated from the totals over three sessions.

Table 4 also shows a highly significant REA for BHR. Interestingly, however, his REA with two-formant stimuli was much smaller and did not reach significance. This finding, which suggested that auditory stimulus complexity may influence the REA, was followed up in Experiment IV.

Actually, the ear advantages were slightly underestimated because one-step contrasts, which were mostly within categories (e.g., 1+2 and 3+4), were included. Of the 21 individual stimulus pairs, 20 showed a positive average ϕ_E. There was a tendency toward larger REAs with increasing separation of the component stimuli on the continuum: the average ϕ_E increased from +0.04 (two-step pairs) to +0.08 (three-step pairs) to +0.11 (four-, five- and six-step pairs), despite the occurrence of uninformative psychoacoustic fusions at the largest separations. Hence, acoustic stimulus disparity may play a role in determining the magnitude of the REA, a question of considerable theoretical importance that deserves further study.

Table 4 (center) also shows the average stimulus dominance (ϕ_D) indices for the individual subjects. These indices express the average dominance of i over j, summed over all $i<j$; or, in other words, the degree of perceptual dominance of lower-frequency F_2 transitions over higher-frequency F_2 transitions (assuming that competition between F_3 transitions plays only a minor role). This average index is rather crude, but it captures some striking individual differences. The overall ϕ_D was positive and highly significant, indicating strong dominance of lower-frequency transitions. However, 2 of the 13 subjects had highly significant negative coefficients.

The ϕ_D indices for the individual stimulus pairs, which were of primary interest, were by no means homogeneous, as was already evident from Figure 1. Only a few pairs were in perceptual equilibrium ($\phi_D \approx 0$), and stimulus dominance effects were considerably stronger than ear dominance effects. The stimulus dominance pattern for a subgroup of 7 of the 13 subjects is illustrated in Figure 2 (filled triangles). This subgroup was selected for reasons of comparison with the results of Experiment II; their data are representative of all 13 subjects, except that the average ϕ_D was somewhat reduced. Discussion of the dominance pattern will be reserved for the General Discussion section following the description of Experiment II.

EXPERIMENT II

The relatively low percentages of psychoacoustic fusions in Experiment I may have been due to random variability in ear dominance or stimulus dominance from trial to trial. Psychoacoustic fusions may occur only when the two syllables receive very nearly equal weights in the hypothetical auditory averaging process; a slight tip of the balance in favor of one stimulus may lead to perceptual dominance of that stimulus. However, when the two syllables in a pair are acoustically combined before they reach the ear, the potential factor of variability in ear dominance is excluded. In addition, auditory averaging may occur at a more peripheral stage and may reduce any variability arising at more central levels. Therefore, this hypothesis predicted an increase in psychoacoustic fusions for mixed stimuli.

A comparison of dichotic and mixed pairs promised to be interesting with respect to the whole "dominance pattern" of individual stimulus combinations. The peripheral interactions coming into play in the mixed mode (acoustic interference, auditory masking) may well lead to an entirely different response pattern than in the dichotic mode. On the other hand, any significant similarities between the two situations will have to be ascribed to common central processing levels.

Method

Subjects. Nine of the thirteen subjects in Experiment I participated, one of them prior to Experiment I. The data of one additional subject were eliminated because they were too noisy.

Materials. The same stimulus tape as in Experiment I was used.

Procedure. The procedure was identical with that of Experiment I except that the output of the two tape recorder channels was mixed electronically and presented binaurally. The intensity was readjusted to about 85 dB SPL. Special care was exercised in equating the intensities of the two channels before they entered the mixer. There was no reversal of channels here.

Results and Discussion

Controls. A comparison of the response distributions for pairs of identical syllables in the dichotic and mixed conditions revealed significant differences for six of the seven syllables. However, the changes consisted primarily in a reduction of the "noise" and an increase in response consistency, so that familiarity and practice were the most likely cause. In view of these changes in the "baseline" scores, it was especially important to compare the response patterns in the two conditions by means of a measure that takes these changes into account. This was achieved by weighting the data as in Experiment I [cf. Eq. (1)], with "channels" replacing "ears." Subsequently, ϕ_D and ϕ_C ("channel dominance") coefficients were calculated [cf. Eqs. (2) and (3)].

While, at the levels used here, intensity differences of a few decibels have little effect in dichotic listening (Speaks and Bissonette, 1975), the mixing procedure was likely to be sensitive to small channel imbalances. The ϕ_C coefficients served as a check on the proper equalization of the two channels prior to mixing. Two subjects indeed showed highly significant ϕ_C coefficients, both in the same single session. This indicated a calibration error, and the data were excluded from further consideration.

Psychoacoustic fusions. Table 5 compares the responses to the four /bæ + gæ/ pairs in the dichotic and mixed conditions for the same seven subjects. Surprisingly, D responses were clearly less frequent in the mixed condition than in the dichotic condition, with B responses making up for most of the difference. This was probably not a practice effect, since BHR--who again participated in five sessions--showed precisely the same decline in psychoacoustic fusions (Table 3, left portion), and a correction for expected D confusions did not eliminate the difference. It may be noted that the data of Halwes (1969) showed a similar reduction in psychoacoustic fusions for mixed syllables.

Since it was conceivable that again the presence of a third formant somehow played a role, BHR once more served as a control subject and listened to mixed two-formant syllables (30 blocks in 3 sessions). The results showed an increase in psychoacoustic fusions with respect to mixed three-formant syllables but a reduction with respect to dichotic two-formant syllables (Table 3). This shows that the reduction was not due to a change in the salience of the third formant.

The stimulus dominance pattern. The overall ϕ_D coefficient was again significant and in favor of the stimuli with the lower numbers on the continuum

TABLE 5: Response percentages for four dichotic /bæ + gæ/ pairs: within-subject comparison of dichotic and mixed conditions (seven subjects).

	Responses					
	Dichotic			Mixed		
Stimulus pair	B	D	G	B	D	G
2 + 6	32.1	32.1	35.8	42.1	19.3	38.6
2 + 7	38.6	32.1	29.3	46.4	17.9	35.7
1 + 6	55.7	16.4	27.9	76.4	2.9	20.7
1 + 7	50.7	15.0	34.3	59.3	5.7	35.0

(lower-frequency F_2 dominance) but slightly reduced in comparison to the dichotic condition (Table 4). Again, there were large individual differences, also from one condition to the other (cf. Table 4).

The stimulus dominance indices for the individual stimulus pairs in the two conditions are compared in Figure 2. The ϕ_D values in Figure 2 represent the dominance of the stimulus held constant in each panel over the stimuli on the abscissa. (Each individual stimulus combination, i + j, may be found twice in Figure 2, once in the panel for i with j on the abscissa, and once in the panel for j with i on the abscissa, with a ϕ_D coefficient of opposite sign. Of course, $\phi_D = 0$ for identical pairs.) It is evident that, with few exceptions, the functions for the mixed condition exhibit the same basic peaks and valleys as those for the dichotic condition. There are some consistent differences as well, primarily in pairs containing stimuli 1 and 2: in the mixed condition, these /bæ/ stimuli showed increased dominance over /gæ/ stimuli (5, 6, 7) but reduced dominance over /dæ/ stimulu (3, 4). The dominance relationship between /dæ/ and /gæ/ stimuli did not change very much.

BHR's data were in excellent agreement with those of the seven subjects. The dominance pattern of BHR's two-formant results was virtually identical to that of his three-formant results, in both the dichotic and mixed conditions, suggesting a negligible role of the third formant apart from its effect on the frequency of psychoacoustic fusions (which were neutral with regard to dominance relationships). Consequently, the differences between the dichotic and mixed conditions were the same for two-formant and three-formant syllables.

GENERAL DISCUSSION: DICHOTIC INTEGRATION

It was noted earlier that a simple "auditory averaging" model--which assumes that a single auditory stimulus, somehow intermediate between the component stimuli, is interpreted phonetically--is somewhat inadequate in explaining the data. It predicts more psychoacoustic fusions than were actually obtained, especially in the mixed condition where auditory averaging should have been perfect, and it cannot account for the effect of stimulus dissimilarity on psychoacoustic fusions (found also by Cutting, 1976). The model may be modified to include random variation in the weights of the averaging process, although the source of the variation is obscure in the mixed condition. Alternatively, one could assume that, in analogy to vision, fusion (auditory averaging) alternates with rivalry (dominance), the probability of rivalry increasing with

FIGURE 2

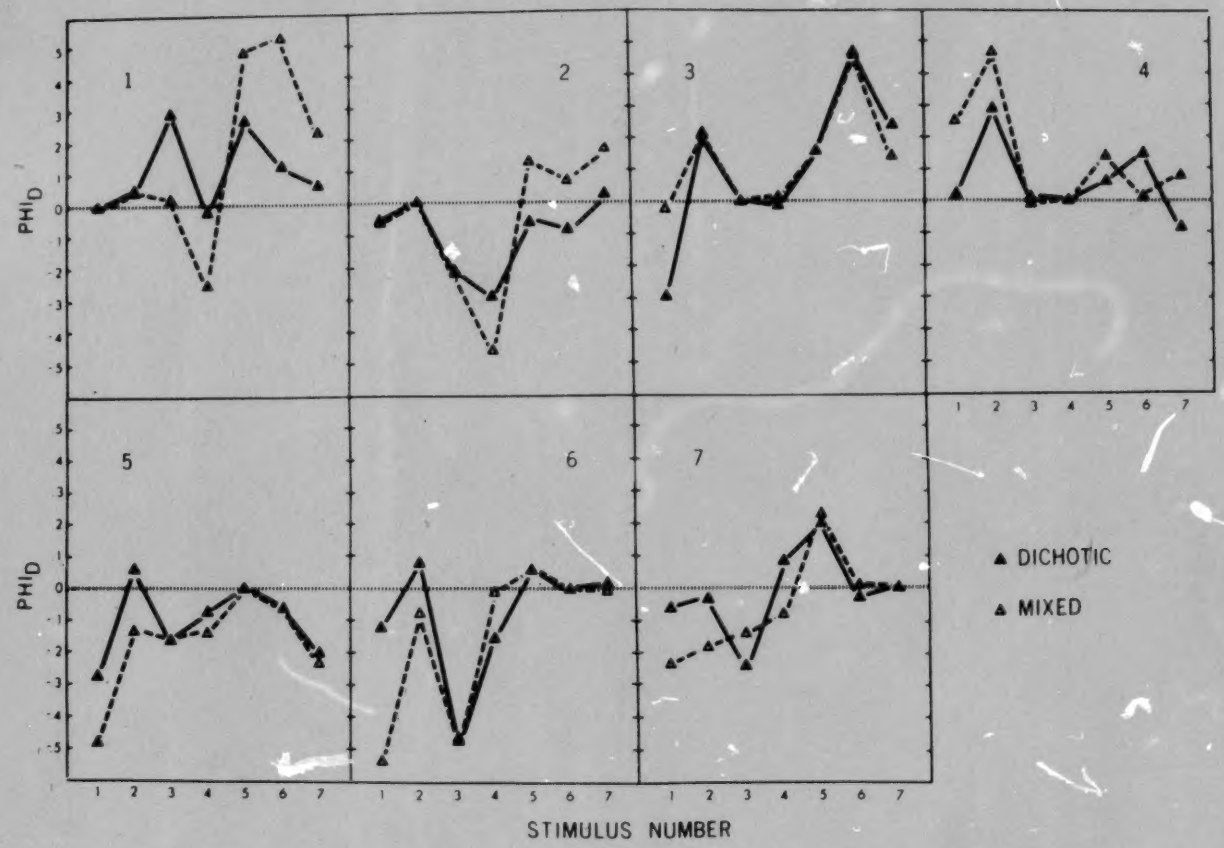


Figure 2: Within-subject comparison of dichotic and mixed conditions: dominance indices (ϕ_D) for each of the seven stimuli (panels 1-7) when paired with the stimuli on the abscissa.

stimulus dissimilarity (Cutting, 1976). While this would account for the pattern of psychoacoustic fusions, the usefulness of a special model for this specific phenomenon is limited. Clearly, psychoacoustic fusions should be explainable by the same principles of interaction as other responses. In other words, an appropriate model should explain the total dominance pattern.

The simple auditory averaging model and Cutting's fusion-rivalry model allow for variable dominance relationships between pairs of stimuli, but only in a form that is related to auditory parameters. For example, consider the dominance function for stimulus 1 in Figure 2: since the starting frequency of the F_2 transition increases monotonically with stimulus number, the dominance function for 1 was expected to be a monotonic function (rising if lower frequencies tend to dominate higher frequencies, and falling if the opposite is true). Because of the possibly special status of straight formants, a smooth curvilinear function would also be reasonable. (BHR's data suggested that the third formant played only a negligible role.) However, there is no straightforward auditory explanation for the abrupt and striking dip of the function at stimulus 4 (that is, for the pair 1+4) and the equally abrupt reversal at stimulus 5 (1+5). Similar observations may be made in several other panels of Figure 2 (for example, panels 3 and 6). The data from the mixed condition weigh especially heavy here. Apparently, then, even when two stimuli are acoustically superimposed and/or perceived as a single syllable, the perceptual mechanism does <u>not</u> treat the composite information simply as the auditory average of its two constituents.

Therefore, we must turn to a different model. The model to be suggested assumes that the acoustic cues of the component stimuli remain independent and largely intact beyond the auditory processing stage, even in mixed syllables, where a stimulus with a rising transition plus one with a falling transition results in a fused stimulus with both a rising <u>and</u> a falling transition. We assume that to this composite information a pattern recognition process is applied that consists in comparing it with "ideal" representations ("prototypes" or "schemata"--cf. Posner, 1969; Rosch, 1975) of the relevant speech sounds in long-term memory. From these ideal representations or prototypes, the one is selected that matches the input most closely.

This process of speech recognition can be conceived as active or as passive (Morton and Broadbent, 1967). The active form is usually referred to as analysis-by-synthesis, pattern matching, or hypothesis testing. The passive form, which is preferred here on heuristic grounds, may be formulated in terms of Morton's "logogen model" (Morton, 1969) or in terms of banks of selectively tuned feature detectors (e.g., Cooper and Nager, 1975). An equivalent but more abstract conception is in terms of a multidimensional perceptual space whose dimensions are the derived auditory characteristics of the relevant set of speech sounds. The relevant response alternatives are located as fixed "ideal points" in this n-dimensional space, while an incoming stimulus generates a point at some location corresponding to its auditory properties. Because synthetic stimuli are acoustically much simpler than real speech (which the prototypes represent), they will be mapped into a subspace of lower dimensionality, for example, a F_2-F_3-transition-frequency plane, in the present case. The distances from the stimulus point to all prototypes are assessed in parallel, and a subsequent decision process selects the prototype with the shortest associated distance as response. A more concrete conceptualization of the calculation of distances is in terms of a "spread of excitation" from the stimulus points,

which leads the prototypes to be activated or to "resonate" in proportion to their distance from the stimulus point.

The model thus comprises three states: (1) Auditory processing, which maps an acoustic stimulus into perceptual space; (2) multicategorical processing, which generates a multicategorical vector of prototype activation values; and (3) a (uni)categorical decision, which selects the response category by determining the largest element in the multicategorical vector. (Stages 2 and 3 constitute what has been traditionally called phonetic processing--Pisoni, 1975; Studdert-Kennedy, in press.)[8]

Random variability may arise at any of the three processing levels: in the representation of the stimulus points in perceptual space ("perceptual noise"--cf. Repp, 1975b), or in the baseline activation levels of the prototypes, or perhaps in the final decision process itself. These details will not concern us further here. The point to be made is that a stochastic pattern recognition model of this sort may provide a useful framework for explaining the speech recognition process, even when applied in an informal (that is, nonnumerical) fashion.

This model should apply to dichotic fusions or mixed stimuli as well as to any single input. Since the location of a stimulus point in the hypothetical perceptual space is determined by its derived auditory characteristics (its "acoustic cues"), and since fused or mixed stimuli contain multiple cues (for example, two different transitions of the same formant), they will lead to two stimulus points in perceptual space. The listener is usually not aware of this fact but only of the perceptual outcome that will be determined by the prototype that reaches the highest level of activation from the simultaneous presence of the two stimulus points.

This assumption predicts the most important feature of the data: the pattern of dominance relationships. The model implies that, of two fused stimuli, that stimulus will dominate that is closer to a prototype in perceptual space. In other words, stimuli close to a category boundary and far from the category prototypes will tend to be dominated by stimuli that are far from a category boundary and close to a prototype. This is what Figure 2 seems to show, on the whole. Stimulus 1, for example, dominates 5 precisely because the latter is ambiguous, whereas it does not dominate 4, which is a good /dæ/; and it dominates 7 only slightly, since 7 is a good /gæ/. Stimulus 2, which is a less

---

[8]This "holistic" model automatically takes into account certain interactions between the processing of different features of a speech sound. An alternative model might postulate that "multicategorical" processing takes place at the auditory level, by means of selectively tuned feature detectors (e.g., Cooper and Nager, 1975) that act as auditory prototypes. This auditory stage would then be followed by a series of feature decisions whose outcomes are finally combined into a response. However, this model would have to explain why the feature detectors are selectively tuned as they are, and it would have to include additional mechanisms for the interaction of different feature decisions. It is worthwhile, therefore, to adopt the holistic model as a working hypothesis, until there is sufficient reason to reject it. We cannot decide between the two models on the basis of the present data because only a single feature is involved.

perfect /bæ/, tends to be dominated by most other stimuli, and so on. The predictions of the model are not confirmed in every detail, but they nevertheless seem to provide the best explanation of the overall pattern.

However, there are other features that the model cannot explain as it stands. Note that stimulus 7 is dominated most strongly by 3, while 1 is dominated most strongly by 4 (Figure 2). In addition, psychoacoustic fusions and the differences between dichotic and mixed pairs need to be accounted for.

Psychoacoustic fusions are explained as follows: if a stimulus in isolation receives 100 percent B responses, this does not mean that only the B prototype has been activated by this stimulus. Because of the hypothetical spread of excitation, all prototypes will be activated to some degree; but if the stimulus is sufficiently close to the B prototype and the noise in the system is not too high, the activation levels of the other prototypes will never exceed the level of the B prototype. However, in dichotic competition the activation resulting from the two stimulus points will be integrated by the prototypes, and since the D prototype is likely to lie somewhere between the B and G prototypes in perceptual space, it will profit most from this integration. If both the /bæ/ and the /gæ/ stimulus in a pair are close to the D boundary, their joint activation of the D prototype may even exceed that of the B and G prototypes. So, for example, 2+7 should yield more D responses than 1+7; and 2+6, more than 1+6, which was in fact obtained (cf. also Cutting, 1976). The component stimuli, 6 and 7, on the other hand, had no differential effect of D responses, which seems to imply that their activation of the D prototype was equal in degree. This is not quite in accord with the model, but it is plausible that differences at higher frequencies have a smaller effect than differences at lower frequencies.

The same reasoning explains why 1 was dominated most strongly by 4, but 7 was dominated most strongly by 3. Clearly, 3 is more likely to activate the B prototype than 4, so that, in the pair 1+3, the B activations will summate and outweigh the D activation due primarily to 3 alone. In 1+4, 4 will contribute less to the activation of the B prototype and D will have a stronger stand against B. The opposite argument applies when 3 and 4 are paired with 7. (These relationships are also predicted by the auditory averaging model.)

The prototype model cannot account for the differences between the dichotic and mixed conditions. Most likely, this difference can be traced back to peripheral auditory masking, which comes into play in the mixed condition. The data suggest that, in mixed syllables, rising transitions (in stimuli 1 and 2) tended to mask (dominate) falling transitions, and relatively flat formants (stimuli 3 and 4) tended to mask rising transitions. The first effect may reflect the "upward spread of masking" familiar from the auditory masking literature, while the second effect may reflect a higher susceptibility to masking of transitions in general, as compared to steady-state formants. The reduction in psychoacoustic fusions in the mixed condition was most likely due to the masking of /gæ/ by /bæ/, so that B responses increased at the expense of D and G responses.

The results pertaining to ear advantages will be discussed after two additional experiments have been reported.

EXPERIMENT III

This brief experiment served to demonstrate what had been based only on introspective evidence in Experiment I, viz. that dichotic fusions are difficult or impossible to discriminate from binaural syllables. In the Introduction, I have referred to the results of several experiments that seemed to show that place contrasts frequently, but not always (in about 60 percent of the cases), sound like a single syllable (Halwes, 1969; Blumstein and Cooper, 1972; Repp, 1976). However, these studies did not differentiate between voiced and voiceless place contrasts (the latter may be less completely fused than the former), and they employed only a single, unambiguous token from each category, so that the frequent ambiguity of dichotic fusions may have assumed the role of a distinctive cue. To test the proposition that binaural and dichotic pairs cannot be distinguished, ambiguity must be made irrelevant. This is at least partially achieved by using syllables from a place continuum, so that at least one of the identical pairs will be ambiguous (stimulus 5, in the present case). The false-alarm rates ("different" responses) for this ambiguous pair should reveal whether the ambiguity cue plays any role.

Method

Subjects. Eight subjects (four men and four women) participated who had not taken part in Experiments I and II. All subjects were right-handed and without hearing trouble, with the exception of one subject who claimed to have a 5-dB hearing loss in the right ear.

Materials. The stimulus tape of Experiment I was used.

Procedure. This discrimination task was appended to Experiment IV, taking up the last 20 minutes of a session. Each subject listened first to one block of 84 syllable pairs (half identical, half nonidentical) and wrote down "1" when he thought a pair consisted of two identical syllables and "2," when it consisted of two different syllables. (To avoid confusion with the stimulus numbers, these responses will be referred to as "same" and "different," respectively.) During the next block of 84 syllable pairs, the subject merely followed the correct responses that had been filled in on the answer sheet. After this feedback trial, another block of judgments followed. The subjects were instructed that there was an equal number of identical and nonidentical pairs, and that ambiguity was not an indication that two different syllables had been presented.

Results and Discussion

As predicted, average performance was very poor, although slightly above chance (56 percent correct). The performance of three individual subjects was significantly above chance (67, 62, and 58 percent correct, respectively). BHR, who participated in four sessions, performed at chance level (51 percent correct), and so did another highly experienced listener who listened informally. The feedback did not improve performance.

A more detailed analysis was conducted in order to find out whether ambiguity played a role and whether accuracy increased with the acoustic dissimilarity of the syllables in a pair. The data are shown in Table 6. The most ambiguous identical pair, 5+5, did not show an increased false-alarm rate, suggesting that ambiguity did not serve as a distinctive cue in this task. On the other

hand, the "hit rate" for nonidentical pairs increased monotonically with the number of steps separating the two syllables in a pair. At the first glance, this seemed to suggest that within-pair acoustic dissimilarity played a role. However, a closer look at the data showed that this was probably not true, and that the result was due to the confounding of acoustic separation with the acoustic characteristics of the component syllables. (Pairs with large separations did not contain any stimuli from the middle of the continuum.) Table 6 shows that both the hit rates for nonidentical pairs and the false-alarm rates for identical pairs were greatly increased when a pair contained stimulus 1, indicating a strong bias to respond "different." Hit rates were also increased for most pairs containing stimuli 2 or 7, relative to the remaining pairs. However, within these groups of pairs (holding one stimulus constant), no clear relation to acoustic dissimilarity could be discerned.

TABLE 6: Percentages of "different" responses to nonidentical pairs ("hits," off-diagonal) and identical pairs ("false alarms," diagonal).

Stimulus number	1	2	3	4	5	6	7
1	61						
2	69	54					
3	75	50	40				
4	75	31	28	26			
5	69	50	22	28	30		
6	75	50	38	34	25	21	
7	81	50	47	53	47	16	19

The most likely explanation of this pattern of results is that the stimuli from the ends of the continuum had some peculiar acoustic properties, perhaps owing to the steep slope of their transitions. This artifact, which may have been due to limitations of the synthesizer or may have been psychoacoustic in nature, was apparently interpreted incorrectly as a relevant cue. The only exception to this interpretation is the very low rates of "different" responses to the pairs 6+7 and 7+7 (Table 6).

Apart from this issue, the data do provide some evidence of better-than-chance performance of some subjects, which remains an astonishing and somewhat puzzling feat. For all practical purposes, however, it may be concluded that dichotic voiced place contrasts are perceived as single syllables.

EXPERIMENT IV

The fourth experiment served three purposes. First, it attempted to demonstrate the ineffectiveness of selective-attention instructions with dichotic fusions. Although Halwes (1969:Experiment 5) found no effect of selective attention in "fused" syllable pairs, a subsequent experiment of his showed a slight effect (Halwes, 1969:Experiment 6). His stimuli actually included all six stop consonants and were called "fused" only because they had the same fundamental frequency. Repp (1973, 1976) has also demonstrated small selective-attention effects for such stimuli. The question here is whether the components of perfectly fused voiced place contrasts can be attended selectively.

The second purpose was a test of the hypothesis suggested by BHR's smaller REA for two-formant stimuli than for three-formant stimuli in Experiment I (Table 4). It may be that stimulus complexity (which in turn may be related to speech-likeness and naturalness) is positively correlated with the REA obtained. For this purpose, two-formant and three-formant pairs were compared in the same design. The role of the third formant in stimulus dominance relationships was also of interest.

The third purpose of Experiment IV was simply to create a more typical test situation, using only one token from each category, in order to find out how serious the problems of stimulus dominance, stimulus heterogeneity, and individual differences actually are in this more "natural" setting. Any such problems encountered should reinforce the methodological suggestions to be made in the final Discussion.

Method

Subjects. The same subjects as in Experiment III participated. However, the data of one subject who did not hear any /gæ/s at all were excluded and replaced by data for BHR as a subject (from the first of four sessions in which he participated).

Materials. The stimuli were three syllables, with or without third formants, from the same place continuum as in the earlier experiments. The /dæ/ was stimulus 4 of Table 1, the /bæ/ had slightly more extreme transitions than stimulus 1 of Table 1 (starting frequency of F_2: 1232 Hz; F_3, if present: 2180 Hz), and the /gæ/ was intermediate between stimuli 6 and 7 (F_2: 2156 Hz; F_3: 2180 Hz).

The experimental tape contained a brief monaural practice list of 30 random syllables (five replications of each of the six stimuli). This was followed by two blocks of 180 dichotic pairs. Each block contained 10 subblocks, each representing a different randomization of 18 dichotic pairs made up from the nine possible combinations of the three syllables with two formants and with three formants, respectively. (Two-formant and three-formant stimuli were never paired with each other.)

Procedure. After trying to identify the practice syllables (and repeating the series if necessary), the subjects listened twice to the experimental tape, that is, to four blocks of 180 dichotic pairs. For two of these blocks, the subjects were instructed to shift their attention to one side, by whatever means they found suitable. It was explained that the syllables actually consisted of two different inputs, and that only the syllables in the designated ear were to be identified. In the remaining two blocks, no selective attention was required, and the subjects simply wrote down what the fused syllables sounded like. The sequence of attention/no-attention conditions and of left-ear and right-ear selective attention was counterbalanced across subjects.

Results

The data were analyzed as in Experiment I. There was a significant overall REA ($\varnothing_E = +0.07$, $p < .01$). Five of the eight subjects showed significant REAs,

one subject a significant LEA.[9] The hypothesis of a difference in REA for two-formant and three-formant syllables was not confirmed. Although individual subjects showed considerable differences, the average ϕ_E indices were identical. BHR even showed a slightly larger REA with two-formant syllables, contrary to the opposite difference in Experiment I, which had given rise to the hypothesis in the first place.

The effect of selective attention was very peculiar: the differences were precisely in the wrong direction. The ϕ_E coefficients were +0.12 for left-ear attention, +0.03 for right-ear attention, and +0.07 for no-attention. The effect was very similar for two-formant and three-formant stimuli. However, no individual subject showed any clear evidence of consistent positive or negative selective attention effects, so that the inverted pattern may have been due to chance. Two subjects showed an inversion of the REA as a function of selective attention but regardless of the ear attended to.

The frequency of psychoacoustic fusions was low (12 percent), as expected with acoustically dissimilar stimuli. This percentage excludes the data of BHR who, as in Experiment I, showed a much higher frequency (35 percent). Quite surprisingly, and contrary to BHR's control results in Experiment I, psychoacoustic fusions were more frequent with three-formant than with two-formant stimuli (15 vs. 8 percent for the seven subjects; 41 vs. 28 percent for BHR).

There was a reliable difference in the stimulus dominance pattern between two-formant and three-formant syllables, which is shown in Table 7 and may be characterized as a reduction in the "strength" of /dæ/ when the third formant was removed. This was already evident in the identification of binaural pairs: the two-formant /dæ/ received only 86 percent correct responses, while the three-formant /dæ/ received 94 percent. (The intelligibility of the other stimuli did not change.) Table 7 shows that, with three formants present, /dæ/ dominated /bæ/ and /gæ/. With two formants, the pattern was reversed. This indicates that an F_3 transition was more important for /dæ/ than for /bæ/ and /gæ/; and it supports the hypothesis, set forth earlier, that a poor representative of a category will be dominated by better examples of other categories. Again, however, there were large individual differences in dominance patterns.

The ϕ_E coefficients for the three individual stimulus pairs (which were similar for two- and three-formant stimuli) are also shown in Table 7. Surprisingly, /dæ + gæ/ pairs did not exhibit an average REA. BHR (who participated in four sessions) even showed a LEA with this pair, but a clear REA with the other two. However, apart from BHR's data, this phenomenon was not reliable for individual subjects who showed large variations in their ear advantages for individual pairs. Both the /dæ + gæ/ anomaly and the high variability are somewhat disconcerting. It will be recalled that Experiment I did not show any comparable effect.

---

[9] This was the subject who claimed to have a 5-dB hearing loss in the right ear. However, it would be quite surprising if this had been the cause of the dichotic asymmetry, considering that channel differences much larger than 5 dB have only little effect on the dichotic ear advantage at the intensities used here (Speaks and Bissonette, 1975).

TABLE 7: Stimulus dominance indices for individual stimulus pairs, and ear dominance indices (averaged over two- and three-formant stimuli). (Note: A positive ϕ_D index indicates dominance of the stimulus named first.)

	/bæ + dæ/	/dæ + gæ/	/bæ + gæ/
three-formant ϕ_D	-0.31	+0.40	+0.45
two-formant ϕ_D	+0.09	-0.31	+0.32
average ϕ_E	+0.14	-0.01	+0.11

GENERAL DISCUSSION: II. MEASURING THE EAR ADVANTAGE

The presence of a significant average REA for dichotic fusions is evidence that, despite the subjective impression of a single syllable, the information from the two ears remains functionally separated until it converges upon the dominant hemisphere. It makes unlikely a low-level auditory mixing mechanism that combines spectrally similar information and routes it to both hemispheres, because such a mechanism would have to be influenced by hemispheric dominance. Rather, it seems that each stimulus first arrives at the contralateral hemisphere, and integration takes place only when the information is recombined after considerable auditory (and perhaps even initial phonetic) processing in each hemisphere, which has been a common assumption in dichotic listening research (Studdert-Kennedy and Shankweiler, 1970). The REA for dichotic fusions challenges an interpretation in terms of spatial location only (Morais and Bertelson, 1973; Morais, 1975). Since only a single stimulus is heard that is localized in the median plane, the hypothesis that stimuli that come from the right are perceived more accurately does not apply.

The subjective phenomenon of fusion (hearing only a single stimulus) probably does arise from a low-level cross-correlational mechanism, but it is apparently separate from, and unrelated to, the subsequent allocation and integration of information. This has two interesting implications: (1) in the limiting case, identical binaural stimuli may also be independently transmitted to their respective contralateral hemispheres and perceptually combined only at a central level; and, more importantly, (2) the identification of less completely fused dichotic stimuli (e.g., voicing contrasts) should be explainable by the same principles as the identification of dichotic fusions, for example, by the prototype model proposed earlier. This view is in basic agreement with the conclusions of Halwes (1969), who found that subjective fusion versus nonfusion was largely irrelevant to the pattern of responses.

It also follows from these conclusions that other types of dichotic contrasts should lend themselves to the one-response, no-attention requirement ("What does it sound like?") whose advantages over the two-response paradigm have already been outlined in the Introduction (cf. Geffner and Dorman, in press, who used this method successfully with four-year-old children). However, what makes voiced place contrasts especially convenient from a methodological standpoint is (1) that the task is "natural" because the listeners are not aware of different inputs to the two ears, (2) that the fused stimuli do not sound strange (as other dichotic contrasts often do) but similar to binaural syllables, (3) that they do not invite selective attention strategies (however ineffective

they may be), and (4) that relatively few responses are given that are ambiguous with respect to ear dominance (psychoacoustic fusions). The last problem can be completely eliminated by simply omitting /bæ + gæ/ pairs from dichotic tests. A dichotic test composed only of /bæ + gæ/ and /dæ + gæ/ pairs, interspersed with binaural controls, should be a useful instrument to try out.

However, such a test still presents some major problems. Foremost among these is the phenomenon of stimulus dominance and the large individual variations connected with it. Extreme dominance of one stimulus in a pair must be prevented; otherwise, this dichotic pair will provide no information about ear dominance. Then, there is the important question of the relationship between stimulus dominance and ear dominance that parallels, but is not identical with, the question of the relationship between performance level and ear dominance in the two-response paradigm (Kuhn, 1973).[10] Finally, there is the question of item homogeneity: Do different dichotic pairs measure the ear advantage to the same degree, even if they have equal stimulus dominance coefficients?

Unlike performance level in the two-response paradigm, which is a global index and cannot be manipulated by the experimenter, stimulus dominance is a characteristic of individual stimulus pairs and can be controlled to a certain degree by manipulating stimulus parameters, as demonstrated in Experiment I. There are two possible ways of making use of this control. One is to try to minimize stimulus dominance and to bring all stimulus pairs as close to equilibrium ($\phi_D = 0$) as possible. Because of individual differences, construction of a single optimal test is out of the question. An appropriate method would be testing under computer control, where, during an initial adaptive phase of testing, the computer keeps track of the responses and adjusts the stimulus parameters to reduce asymmetries. Such a procedure is worth exploring but has some drawbacks: it does not guard against drifts of stimulus dominance during the actual testing phase, and it requires sophisticated equipment and therefore is of little value outside the laboratory. The other alternative is to construct a test containing a variety of stimuli, so that the individual pairs span a wide range of stimulus dominance relationships (as in Experiment I). In order to derive a valid measure of ear dominance, in this case, the nature of the relationship between stimulus dominance and ear dominance must be known. Since it is reasonable to expect that ear dominance will be maximal when stimulus dominance is minimal, a global ϕ_E index obtained from summed response frequencies (as in Experiment I) or from averaged ear dominance coefficients for individual pairs will underestimate the "true" ear advantage and will not be comparable from individual to individual, because of different individual stimulus dominance patterns. A method for inferring the true ear advantage is needed.

The situation is formally analogous to that in signal detection. Ear dominance represents "sensitivity" and stimulus dominance represents "bias." When there is extreme bias ($\phi_D = \pm 1$), sensitivity cannot be determined ($\phi_E = 0$). When sensitivity is optimal ($\phi_E = \pm 1$), there cannot be any bias ($\phi_D = 0$).

---

[10] The question of performance level also arises in the present paradigm, in the form of confusions. As long as the confusions are not too numerous, however, their impact is negligible because of the weighting procedure employed [Eq. (1)]. There are some individuals, however, who seem to be unable to give consistent identification responses to the synthetic syllables used here.

Between these extremes, the two tendencies mutually constrain each other. For example, when $T_{(i)}/N = 0.8$ ($\phi_D = +0.75$), it can easily be shown that T_{RE}/N is restricted to the range between 0.3 and 0.7 (ϕ_E between ± 0.5); and ϕ_E constrains ϕ_D in a similar fashion. In order to apply the methods of signal detection theory, one event (for example, responding i when i-j is presented, with i in the right ear) may be arbitrarily chosen to represent "hits," and another event (responding i when j-i is presented, with i in the left ear), "false alarms." However, the crucial requirement is that sensitivity (namely, the "true" ear advantage) be independent of the bias (stimulus dominance). Since stimulus dominance is varied by changing the characteristics of the stimuli (rather than by manipulating the listeners' criteria), it is an important empirical question whether all items are homogeneous (in the test-theoretical sense) and measure the same kind of ear advantage, so that all stimulus pairs can be represented as points on the same single receiver-operating-characteristic function.

The results of the present experiments create some doubts about whether the homogeneity assumption will be tenable. When plotted as "hits" versus "false alarms," the stimulus pairs of Experiment I exhibited considerable scatter, perhaps owing to the high individual variability in the data. There was also a tendency for ϕ_E to increase with the acoustic dissimilarity of the component stimuli in a dichotic pair. At the same time, there was no negative correlation between ϕ_E and $|\phi_D|$ ($r = +0.04$), so that an increase in ϕ_E could not be explained by a simultaneous decrease of dominance asymmetries. In Experiment IV, one of the three stimulus pairs showed no REA. Again, this was not related to stimulus dominance (cf. Table 7). As a result, no monotonic receiver-operating characteristic function will fit these data well. Further research will be required to determine the reliability of the present findings. It may be useful to compare variations in stimulus dominance produced by varying stimulus parameters with similar variations introduced by other means, such as adaptation (Cooper, 1974; Miller, 1975).

A more explicit model of dichotic interaction would also contribute to the solution of this methodological problem. In mathematical terms, stimulus dominance (bias) and ear dominance (sensitivity) mutually constrain each other. However, in the actual processing chain, the constraint may well be unidirectional, since it is highly likely that the two asymmetries arise at different stages in processing. Since stimulus dominance effects were more pronounced than ear dominance effects but did not correlate with the latter, the present data suggest that the cause of ear dominance precedes the cause of stimulus dominance in the processing hierarchy. This is in agreement with the hypothesis that ascribes ear dominance to transcallosal transmission loss but stimulus dominance to subsequent integration of information in the dominant hemisphere.

REFERENCES

Blumstein, S. and W. Cooper. (1972) Identification versus discrimination of distinctive features in speech perception. Quart. J. Exp. Psychol. 24, 207-214.

Cooper, W. E. (1974) Adaptation of phonetic feature analyzers for place of articulation. J. Acoust. Soc. Am. 56, 617-627.

Cooper, W. E. and R. M. Nager. (1975) Perceptuo-motor adaptation to speech: An analysis of bisyllabic utterances and a neural model. J. Acoust. Soc. Am. 58, 256-265.

Cutting, J. E. (1972) A preliminary report on six fusions in auditory research. Haskins Laboratories Status Report on Speech Research SR-31/32, 93-107.

Cutting, J. E. (1976) Auditory and linguistic processes in speech perception: Inferences from six fusions in dichotic listening. Psychol. Rev. 83, 114-140. [Also in Haskins Laboratories Status Report on Speech Research SR-44 (1975), 37-73.]

Geffner, D. S. and M. F. Dorman. (in press) Hemispheric specialization for speech perception in four-year-old children from low and middle socioeconomic classes. Cortex. [Also in Haskins Laboratories Status Report on Speech Research SR-42/43 (1975), 241-245.]

Haggard, M. (1975) The terrible truth about the masking of monosyllables. Speech Perception, Report on Speech Research in Progress (Psychology Department, The Queen's University of Belfast) Series 2, no. 4, 21-30.

Halwes, T. G. (1969) Effects of dichotic fusion on the perception of speech. Unpublished Ph.D. dissertation, University of Minnesota.

Harris, K. S., H. S. Hoffman, A. M. Liberman, P. C. Delattre, and F. S. Cooper. (1958) Effect of third-formant transitions on the perception of the voiced stop consonants. J. Acoust. Soc. Am. 30, 122-126.

Hoffman, H. S. (1958) Study of some cues in the perception of the voiced stop consonants. J. Acoust. Soc. Am. 30, 1035-1041.

Kuhn, G. M. (1973) The phi coefficient as an index of ear differences in dichotic listening. Cortex 9, 447-457.

Miller, J. L. (1975) Properties of feature detectors for speech: Evidence from the effects of selective adaptation on dichotic listening. Percept. Psychophys. 18, 389-397.

Morais, J. (1975) The effects of ventriloquism on the right-side advantage for verbal material. Cognition 3, 127-139.

Morais, J. and P. Bertelson. (1973) Laterality effects in diotic listening. Perception 2, 107-111.

Morton, J. (1969) Interaction of information in word recognition. Psychol. Rev. 76, 165-178.

Morton, J. and D. E. Broadbent. (1967) Passive versus active recognition models, or is your homunculus really necessary? In Models for the Perception of Speech and Visual Form, ed. by W. Wathen-Dunn. (Cambridge, Mass.: MIT Press).

Pisoni, D. B. (1971) On the nature of categorical perception of speech sounds. Unpublished Ph.D. dissertation, University of Michigan.

Pisoni, D. B. (1975) Dichotic listening and processing phonetic features. In Cognitive Theory: Volume 1, ed. by F. Restle, R. M. Shiffrin, N. J. Castellan, H. Lindman, and D. B. Pisoni. (Hillsdale, N. J.: Lawrence Erlbaum Assoc.).

Pisoni, D. B. and S. D. McNabb. (1974) Dichotic interactions of speech sounds and phonetic feature processing. Brain Lang. 1, 351-362.

Posner, M. I. (1969) Abstraction and the process of recognition. In The Psychology of Learning and Motivation. Advances in Research and Theory: Volume 3, ed. by G. H. Bower and J. T. Spence. (New York: Academic).

Repp, B. H. (1973) Dichotic forward and backward masking of CV syllables. Unpublished Ph.D. dissertation, University of Chicago.

Repp, B. H. (1975a) Dichotic forward and backward "masking" between CV syllables. J. Acoust. Soc. Am. 57, 483-496.

Repp, B. H. (1975b) Distinctive features, dichotic competition, and the encoding of stop consonants. Percept. Psychophys. 17, 231-240.

Repp, B. H. (1976) Effects of fundamental frequency contrast on identification and discrimination of dichotic CV syllables at various temporal delays. Mem. Cog. 4, 75-90.

Rosch, E. (1975) The nature of mental codes for color categories. J. Exp. Psychol.: Human Perception and Performance 1, 303-322.

Shankweiler, D. and M. Studdert-Kennedy. (1967a) Identification of consonants and vowels presented to left and right ears. Quart. J. Exp. Psychol. 19, 59-63.

Shankweiler, D. and M. Studdert-Kennedy. (1967b). An analysis of perceptual confusions in identification of dichotically presented CVC syllables. Haskins Laboratories Status Report on Speech Research SR-10, 63-73.

Shankweiler, D. and M. Studdert-Kennedy. (1975) A continuum of lateralization for speech perception? Brain Lang. 2, 212-225.

Speaks, C. and L. Bissonette. (1975) Interaural-intensive differences and dichotic listening. J. Acoust. Soc. Am. 58, 893-898.

Studdert-Kennedy, M. (1975) Two questions. Brain Lang. 2, 123-130.

Studdert-Kennedy, M. (in press) Speech perception. In Contemporary Issues in Experimental Phonetics, ed. by N. J. Lass. (New York: Academic).

Studdert-Kennedy, M. and D. Shankweiler. (1970) Hemispheric specialization for speech perception. J. Acoust. Soc. Am. 48, 579-594.

Discrimination of Dichotic Fusions

Bruno H. Repp*

ABSTRACT

The discriminability of dichotic fusions (dichotic voiced stop-consonant-plus-vowel syllables from the "place continuum" /bæ/-/dæ/-/gæ/) was assessed in an AXB paradigm by presenting stimuli composed of a variable stimulus in one ear and a constant stimulus (either /bæ/ or /gæ/) in the other ear. In a control condition, the variable stimuli were presented without the constant stimulus. On the "categorical-perception" assumption that syllables are discriminated only as well as their labels, dichotic discrimination performance was predicted to be poor and without the typical peaks and troughs observed in single-channel discrimination. However, the obtained discrimination functions showed basically the same peaks and troughs as the single-channel functions, regardless of the nature of the constant stimulus; only performance was lower. A second experiment, employing three variants of the dichotic discrimination task, ruled out selective attention to one channel as an explanation. The results strongly suggest that the discrimination of speech sounds is not based on their phonetic labels but on lower-level codes whose discrete elements represent the proximities of the stimuli to several fixed "prototypes" ("multicategorical vectors"). Dichotic integration is assumed to precede discrimination and to consist of a weighted averaging of the multicategorical vectors of the component stimuli. (The weights represent ear dominance effects, which tended to favor the right ear but were not very consistent in the present discrimination tasks.)

INTRODUCTION

Many recent studies of dichotic listening have employed as stimuli the six stop consonants followed by a constant vowel (e.g., /ba/, /da/, /ga/, /pa/, /ta/, /ka/). It has been known for some time that some of the dichotic contrasts made up from these stimuli tend to fuse and sound like a single syllable

---

*Also University of Connecticut Health Center, Farmington.

Acknowledgment: This research would not have been possible without the generous hospitality of Haskins Laboratories and its director, Alvin Liberman. I am deeply grateful to A. M. Liberman for his helpful comments and his interest in this research. The author was supported by NIH grant T22 DE00202 to the University of Connecticut Health Center.

[HASKINS LABORATORIES: Status Report on Speech Research SR-45/46 (1976)]

(Halwes, 1969; Repp, 1976a; Cutting, 1976). In a recent paper, Repp (1976b) demonstrated that dichotic pairs of precisely aligned synthetic syllables differing only in the initial formant transitions (/bæ/, /dæ/, /gæ/) are virtually indistinguishable from binaural syllables; in other words, they fuse perfectly and sound like single (binaural) syllables. Repp (1976b) presented detailed identification data for such dichotic fusions, and he also showed that the characteristic right-ear advantage is obtained with these stimuli and that paying selective attention to one ear has little effect on the responses.

The present studies investigated the discrimination of these dichotic fusions. The principal question was: Is the perception of dichotic fusions categorical? It is well-known that single syllables that differ only in their formant transitions (i.e., syllables from a "place continuum") are discriminated very poorly as long as they fall within the same category (Liberman, Harris, Hoffman, and Griffith, 1957; Eimas, 1963; Pisoni, 1971). Discrimination performance can be fairly accurately predicted from knowledge of the labeling function, assuming that discrimination relies solely on the phonetic categories assigned to the stimuli. Since single syllables are perceived in this categorical fashion, and since dichotic fusions sound like single syllables, it was only reasonable to expect that their perception would likewise be categorical, so that discrimination performance could be accurately predicted from identification data for the same fusions. However, the possibility remained that, despite subjective fusion, information from the individual channels might be accessible to some degree in a discrimination task; in this case, performance should be better than predicted.

The task selected was termed "one-ear discrimination." It required the listener to discriminate between two dichotic fusions that differed only in the component presented to one ear (the variable stimulus) but not in the other component (the constant stimulus--cf. Figure 2). Of course, the subjects were not aware of the separate components but heard only single, fused syllables. In a control condition, the variable stimuli were presented by themselves, without the constant stimulus. By comparing the results of this single-channel control with those of dichotic one-ear discrimination, the effect of the constant stimuli could be ascertained. "Categorical-perception" predictions were derived from the identification data in Repp (1976b).

A secondary question concerned the dichotic right-ear advantage (REA). Since the variable stimulus could occur either in the left or the right ear, one-ear discrimination performance was expected to be higher when it was in the right ear. The magnitude of the REA could actually be predicted from the identification data, and the one-ear discrimination task seemed interesting as a possible alternative to identification tasks in assessing ear advantages. On the other hand, the REA might turn out to be either larger or smaller than predicted. The first outcome would suggest that the discrimination task is a more sensitive indicator of ear asymmetries than the identification task, while the second outcome would suggest that the listeners base their discriminations on stimulus codes that are less lateralized or bilaterally represented. Both outcomes would be in disagreement with the assumptions of categorical perception.

Before the experiments are discussed, two remarks on methodology are in order.

An AXB discrimination paradigm was used in all the present studies: three successive stimuli were presented, and the listener had to decide whether the second stimulus was equal to the first (AAB or same-different configuration) or to the third (ABB or different-same configuration). This paradigm has been rarely used in the past, although it seems to combine the advantages of the more popular ABX and 4IAX paradigms (Pisoni, 1971; Pisoni and Lazarus, 1974). Pisoni has demonstrated that the 4IAX paradigm, which consists in judging which of two successive pairs of stimuli contain a difference, leads to higher performance than the ABX paradigm, presumably because of the possibility of a "second-order" comparison between subjective differences. The AXB paradigm also allows such second-order comparisons (of the A-X difference with the X-B difference), since the two identical stimuli never straddle the odd one (as is the case in the ABA configuration of the ABX paradigm). Thus, AXB may well be as sensitive as 4IAX, but it is as economical as ABX, since only three stimuli are presented in a trial.

In the tasks described here, it is important that the dichotic syllables are exactly simultaneous. Even very small asynchronies may lead to changes in the subjective location of successive fused stimuli (henceforth referred to as "location shifts"), which will aid discrimination and confound the results. Cherry and Sayers (1956) and, more recently, Young, Parker, and Carhart (1975) have shown that the discrimination threshold for temporal asynchronies between binaural speech sounds is as low as 0.02 to 0.03 msec, which sets the upper limit for the permissible error in the present studies. This precision is not achieved by standard procedures for recording dichotic tapes, a fact that was fully realized only after the present experiments (Experiments IA and IIA) had been conducted. Therefore, both experiments were replicated after a procedure for more precise syllable alignment had been devised, and the original studies will be described together with their replications (Experiments IB and IIB). With the exception of one part of Experiment IIA, which showed evidence of artifacts, the replications confirmed the original data.

EXPERIMENTS IA AND IB

Method

Subjects. There were seven subjects in Experiment IA and seven different subjects in the replication, Experiment IB. All were paid volunteers, right-handed, and relatively inexperienced listeners. The subjects of Experiment IA had previously participated in an identification task using the same stimuli (Repp, 1976b: Experiment I). The data of one additional subject in each study were excluded because they were at chance level.

Stimuli. The stimuli were seven syllables from a "place continuum" (Pisoni, 1971), ranging perceptually from /bæ/ to /dæ/ to /gæ/. They were produced on the Haskins Laboratories parallel resonance synthesizer. All syllables had the same duration (280 msec), a constant fundamental frequency (114 Hz), a voice onset time (VOT) of -15 msec (i.e., prevoicing), 45-msec linear transitions, and no bursts but an abrupt onset of energy following the prevoicing. They differed only in the onset frequencies of the second-formant (F_2) and third-formant (F_3) transitions, which are shown in Table 1.

TABLE 1: Starting frequencies (in Hz) of second-formant (F_2) and third-formant (F_3) transitions of the seven stimuli.

Stimulus No.	F_2	F_3
1	1312	2348
2	1456	2694
3	1620	3026
4	1772	3026
5	1920	2694
6	2078	2348
7	2234	2018
/æ/	1620	2862

Dichotic pairs were constructed using the pulse code modulation (PCM) system at Haskins Laboratories. This procedure involved digital sampling of the synthesizer output with a standard sampling rate of 8,000/sec in Experiment IA, resulting in a random sampling error not exceeding 0.125 msec, which remained fixed for each individual stimulus. In addition, because the smallest accessible unit was two samples, the onset of a stimulus could be in an even or in an odd sample, so that the onsets of two dichotic syllables could be off by ±0.125 msec. (This was probably the more important factor.) In Experiment IB, all syllables were redigitized until they all started in an odd sample, which eliminated the onset asynchronies. In addition, a faster sampling rate was used (20,000/sec), which reduced the random error to below 0.05 msec. Furthermore, a magnified section of the steady-state vowel of each stimulus was displayed on a storage oscilloscope and compared to a standard waveform selected from one of the stimuli. Poor matches were rejected, and the stimuli were redigitized until their waveforms matched the standard quite well. This procedure reduced the random error to at least half its magnitude and, thus, below the detection threshold for "location shifts."[1]

In Experiment IB, the following characteristics of the stimuli were inadvertently changed: overall duration was reduced to 196 msec, prevoicing to 10 msec, and transition duration to 38 msec.

The experimental tape of Experiment IA contained first a random series of 44 AXB triads of single syllables (i.e., in one channel only). Only the six "one-step" (1 vs. 2, 2 vs. 3, etc.) and the five "two-step" (1 vs. 3, 2 vs. 4, etc.) discriminations were included, in each of the four possible AXB configurations (AAB, ABB, BBA, BAA). This was followed by a series of 88 dichotic triads in which the same (variable) stimuli in one ear were combined with the constant stimulus 1 (/bæ/) in the other ear. The variable stimulus could occur either in the left or the right ear. Another similar series of 88 dichotic triads followed in which the constant stimulus was 7 (/gæ/). Finally, there was another series of 44 single-channel triads. The interstimulus interval was 1 sec and the intertriad interval 3 sec.

[1] It should be noted that neither the author nor any of the other subjects reported any location shifts in Experiment IA. Nevertheless, the replication seemed an appropriate cautionary measure.

The experimental tape of Experiment IB differed from that of Experiment IA in that the two constant stimuli, 1 and 7, were not blocked but randomized, so that the dichotic triads constituted a single block of 176 trials.

Procedure. The subjects were tested in small groups, usually joined by the experimenter, in a single session lasting approximately 90 minutes. Playback was from an Ampex AB-500 tape recorder through an amplifier to Telephonics TDH-39 earphones. Playback intensity was adjusted and monitored on a Hewlett-Packard voltmeter, and special care was taken to equalize the intensities of the two channels at about 85 dB SPL (peak deflections for individual syllables), which was the intensity used in the earlier identification study.

Each subject listened to the experimental tape twice. The earphone channels were interchanged electronically before the second run. The single-channel trials were presented binaurally in Experiment IA but monaurally in Experiment IB. The AXB paradigm was explained in detail: X was described as a variable stimulus that could be equal either to A or to B, the latter two always being different from each other. Correspondingly, the subjects were asked to write down A or B as their responses and to guess when uncertain. The two configurations $A = X \neq B$ (same-different) and $A \neq X = B$ (different-same), were pointed out and appeared in this form as a reminder on the answer sheets. This was intended to guide the subjects to a processing strategy similar to that in a 4IAX paradigm. The subjects were not informed about the dichotic nature of the stimuli until after the experiment.

In summary, Experiment IB differed from Experiment IA by (1) more precise stimulus synchronization, (2) shorter stimulus and transition durations, (3) monaural instead of binaural presentation of single-channel trials, and (4) random instead of blocked sequences of the two constant stimuli in dichotic trials. However, none of these changes was expected to have any great effect, and Experiments IA and IB were expected to agree in their main results.

Results

Single-channel discrimination. The average single-channel discrimination performance in the two experiments is shown in the left-hand panels of Figure 1. The upper panel also shows the functions predicted from the identification data (Repp, 1976b: Experiment I), assuming perfect categorical perception and absence of sequential effects. The prediction formula is the same as in the ABX and 4IAX paradigms (Pollack and Pisoni, 1971).

The discrimination functions show the characteristic peaks and troughs of categorical perception (Pisoni, 1971). They are more pronounced in Experiment IB than in Experiment IA, indicating that the syllables in the replication study were labeled more consistently. There are some deviations from the predictions in Experiment IA. Most of these can be explained by a shift in the labeling of stimulus 5 toward G, relative to the earlier identification experiment. (There, stimulus 5 had been the only truly ambiguous syllable, and it had received somewhat more D than G responses.) The functions in the lower panel indicate that stimulus 5 was consistently labeled G in Experiment IB. One deviation from the predictions that cannot be explained by a shift in labeling responses is the better-than-chance discrimination of 1 from 2; both syllables were perceived as B in the identification study, and other features of the discrimination data

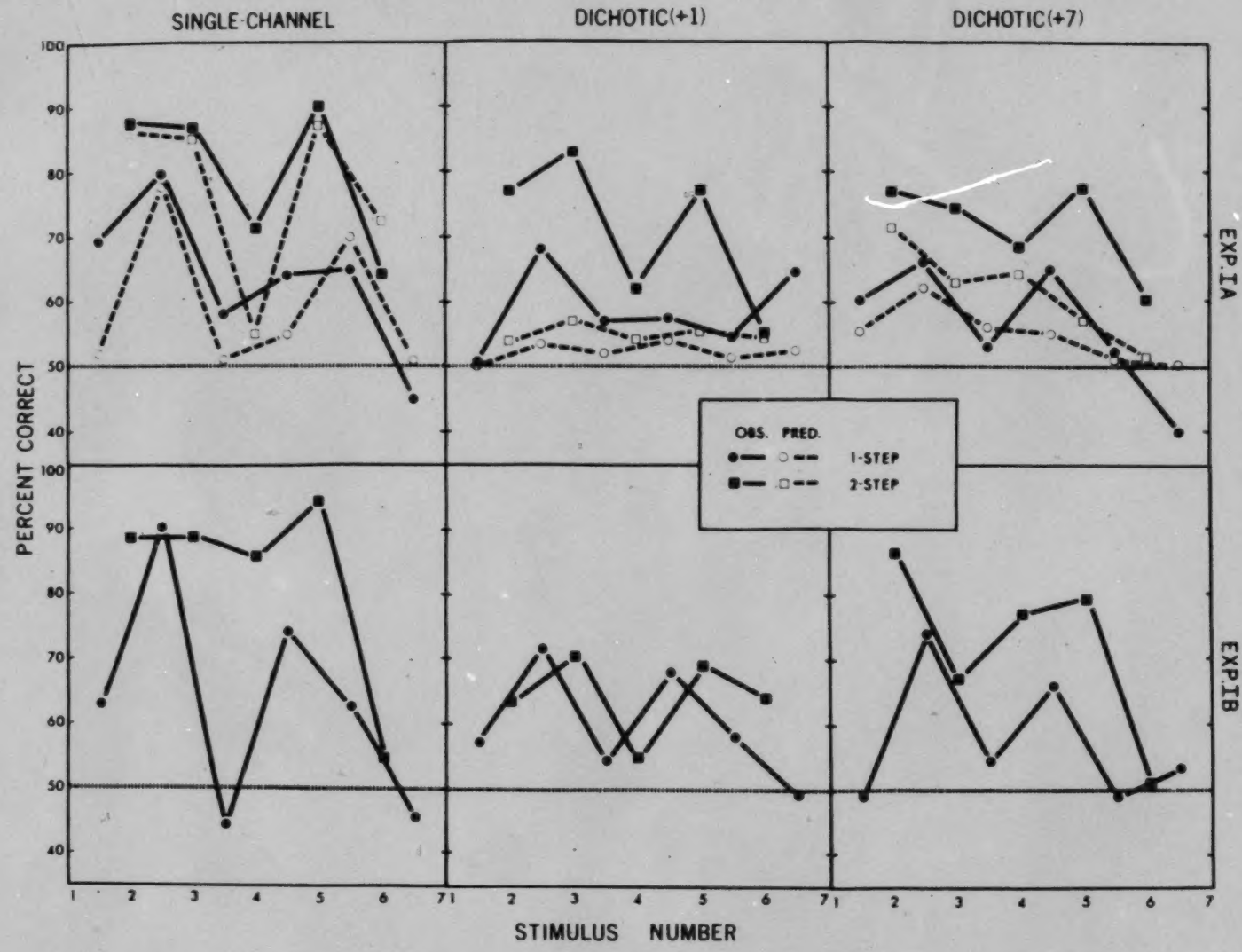


Figure 1: Predicted and obtained one-step and two-step discrimination functions in single-channel and dichotic (one-ear) conditions of Experiments IA and IB. The data points are drawn halfway between the stimuli (abscissa) that were discriminated.

suggest the same for the present studies. This seems to be an instance of true within-category discriminability. On the whole, however, the data are in good agreement with the categorical-perception assumption that the stimuli were discriminated little better than by their labels alone.

One-ear discrimination. The middle and right-hand panels of Figure 1 illustrate the predicted and obtained one-ear discrimination results with 1 and 7, respectively, as constant stimuli. It is obvious that the obtained discrimination functions diverged dramatically from the predicted functions. Performance was expected to be quite poor, especially in the +1 condition (due to stronger perceptual dominance of 1 than 7 in dichotic competition, as observed in the identification study), and no pronounced peaks and troughs in the discrimination functions were predicted (owing to generally inconsistent labeling of dichotic fusions). The obtained functions, on the other hand, did show clear peaks and troughs, and performance was generally much higher than expected.

This was equally true for both experiments, which demonstrates that the results in Experiment IA were not due to artifactual "location shifts." (In any case, no such shifts were heard during the experiment.) An analysis of variance showed no significant difference in overall performance level between the two experiments, nor was there any difference between the overall effects of the two constant stimuli.[2] Those differences that did exist between the two sets of data were probably due to intersubject variability, blocked versus random constant stimuli, and perhaps the changes in acoustic stimulus structure.

A comparison of the dichotic with the single-channel discrimination functions in Figure 1 shows that performance was lower in the dichotic condition but that the pattern remained basically the same. Despite considerable variation in detail, the location of the major peaks and troughs did not change as a function of the constant stimulus in the other ear. The only shift may be seen in the +7 condition of Experiment IB, where the first valley of the two-step function has shifted to the left, that is, away from the constant stimulus.[3]

Ear advantages. The seven subjects in Experiment IA showed only a very small and nonsignificant average REA (ϕ = +0.02; cf. Kuhn, 1973). In the +7 condition, there was actually a small left-ear advantage (LEA), while, in the +1 condition, there was a somewhat larger REA (ϕ = +0.06; $\chi^2(1) = 3.8$; $p = .05$). Although the same seven subjects had shown an average REA of ϕ_E = +0.08 in the

---

[2]The relatively poor two-step discrimination in the +1 condition of Experiment IB was not tested for significance and remains unexplained.

[3]Note that this is the opposite effect from that of adaptation where a migration of peaks and valleys toward the adapting stimulus occurs (Cooper, 1974). Apparently, little adaptation took place in the blocked conditions of Experiment IA. (The possibility of such adaptation effects had prompted the randomization of constant stimuli in the replication study.) The author participated as an additional subject in four sessions of each experiment. His data generally confirmed the results of the less experienced listeners, except that he showed more pronounced migrations of the discrimination peaks away from the constant stimulus end of the continuum. An explanation for this deviant result will be suggested in Footnote 8.

identification task (see Repp, 1976b, about calculation of ϕ_E), the REA in discrimination was predicted to be smaller, so that the obtained average REA was of the expected magnitude. However, there was no agreement with the predictions at a more detailed level, and there was no relationship between the ear advantages in the identification and discrimination tasks. No individual ϕ coefficient reached significance.

In Experiment IB, there was no average ear asymmetry at all, and there was no difference between the +1 and +7 conditions. However, two subjects showed significant individual ear asymmetries (one REA and one LEA).[4]

A joint analysis of variance of the two experiments showed no significant ear advantage. The triple interaction that reflects the REA in the +1 condition of Experiment IA was only marginally significant.

Discussion

The results present an interesting paradox: the perception of dichotic fusions was both categorical and noncategorical. It was categorical because the discrimination functions showed peaks and troughs. At the same time, it was noncategorical because performance was much better than predicted from the identification data. It should be noted that, contrary to single-channel discrimination, no peaks and troughs were predicted for dichotic discrimination because of the absence of clear categories in the identification of dichotic fusions due to the relative dominance of the constant stimulus (see Repp, 1976b: Figure 1). In a sense, all dichotic fusions with a given constant stimulus were within a single, ill-defined category; hence, the poor expected performance.

The discrepancy between predicted and obtained data is evidence that the subjects did not base their discriminations on the labels assigned to the dichotic fusions. What, then, formed the basis of their responses? One obvious possibility, suggested by the general coincidence of the discrimination peaks in the single-channel and dichotic conditions, is that the subjects had access to the information from the separate channels prior to its fusion and integration. Under this hypothesis, they were simply discriminating the variable stimuli and ignored the constant stimuli which only had the effect of noise and led to a generally lower level of performance. In order to test this hypothesis, two new discrimination tasks were devised that are relevant to the question of channel accessibility. Because of the unclear results with respect to the REA, it also seemed desirable to obtain further data on one-ear discrimination, so that Experiments IIA and IIB contained three different discrimination tasks.

EXPERIMENTS IIA AND IIB

The three discrimination tasks are illustrated in Figure 2. The first task was one-ear discrimination, as in Experiments IA and IB. The second task was termed "reversal discrimination." It consisted in telling apart two dichotic

---

[4] The author, who had shown a reliable REA in the identification task, exhibited only a small and nonsignificant REA in Experiment IA (ϕ = +0.02) but a much larger and significant effect in Experiment IB (ϕ = +0.10; $\chi^2(1) = 12.8$, $p < .001$).

fusions made up of the same components, the only difference being the channel (ear) assignment of these components. The third task was a combination of the other two and was called "crossover discrimination." It consisted in discriminating two dichotic fusions that differed in only one component, which, however, "crossed over" to the opposite ear. In other words, crossover discrimination was a "one-ear" discrimination with an additional channel reversal.

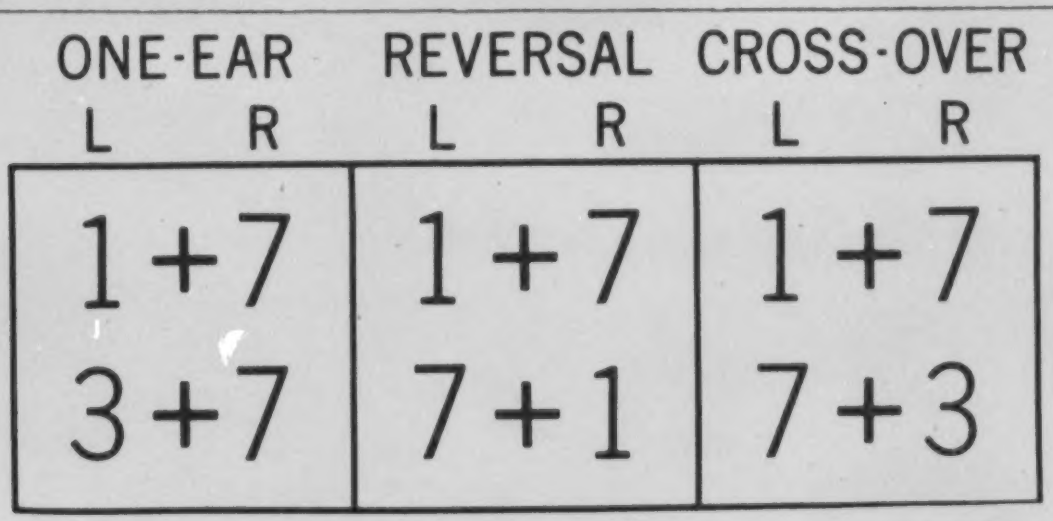


Figure 2: Three discrimination tasks with fused syllables. The numbers represent individual stimuli.

Both new tasks address the channel-accessibility hypothesis. On the categorical-perception assumption that it is the labels that are discriminated, performance in reversal discrimination should be close to chance. In fact, in the absence of any ear asymmetry, reversal discrimination should be impossible. Performance should improve in proportion to the ear advantage (regardless of its direction), but since ear advantages for dichotic fusions are generally small (Repp, 1976b), the expected level of accuracy remained very low. On the other hand, the subjects should be much more successful if they had access to the separate channels, since each channel contains a discriminable difference.

A similar argument may be made for crossover discrimination. On categorical-perception assumptions, crossover discrimination should be as easy (or as difficult) as one-ear discrimination of the same stimuli, except for small differences due to ear asymmetries. However, if the listeners had access to the individual channels, performance should be considerably higher in crossover discrimination. Not only are there discriminable differences in both channels (as opposed to one channel in one-ear discrimination), but these differences are also typically easier to detect than those in the variable channel of one-ear discrimination (cf. Figure 2).

Method

Subjects. There were nine subjects (one left-handed) in Experiment IIA and ten subjects (two left-handed) in Experiment IIB; three of these subjects took part in both experiments.

Materials. The syllables were those of Table 1, with one additional syllable from the lower (/bæ/) end of the continuum; it was called stimulus 0 and had transitions starting at 1155 Hz (F_2) and 2018 Hz (F_3). The recording procedures of Experiments IIA and IIB were identical to those of Experiments IA and IB, respectively.

The experimental tapes contained first a series of 64 triads of single syllables, which were presented monaurally in both experiments. The series contained the four ABX configurations of each of 16 stimulus discriminations: the four two-step discriminations, 1 vs. 3, 2 vs. 4, 3 vs. 5, and 4 vs. 6, and all discriminations of stimuli 0 and 7, respectively, from stimuli 1 through 6. This series was followed by a completely randomized series of 176 dichotic triads comprising 64 one-ear trials, 64 crossover trials, and 48 reversal trials. The one-ear and crossover triads represented the four two-step discriminations with either 0 (/bæ/) or 7 (/gæ/) as the constant stimulus, in all possible AXB and channel configurations. The reversal triads consisted of the dichotic combinations of 0 and 7, respectively, with stimuli 1 through 6, in all AXB configurations.

Procedure. Each subject listened to the tape twice, with a pause in between during which the earphone channels were reversed. Otherwise, the procedure was identical to that in the previous experiments.

Results

Single-channel discrimination. The overall accuracy of monaural discrimination was the same in the two experiments (Experiment IIA: 81.6 percent correct; Experiment IIB: 81.3 percent correct). A more detailed breakdown of the results is shown in the left-hand portions of Tables 2 and 3. Obviously, stimuli 1 and 2 were difficult to discriminate from 0, and 5 and 6 were difficult to discriminate from 7; these stimuli fell within the B and G categories, respectively. Table 2 shows that discrimination from 0 became relatively easier and discrimination from 7 became relatively more difficult in Experiment IIB, both within and between categories. The reason for this interaction is not clear.

Reversal discrimination. The reversal discrimination results are shown in Table 2. In the data of Experiment II, at least three stimulus combinations can be discerned for which artifactual location shifts apparently provided a valid cue (underlined in Table 2), although performance did not exceed 75 percent correct even in those pairs. However, there was surprisingly little change in overall accuracy from Experiment IIA to Experiment IIB; in fact, performance improved for six of the stimulus combinations. This suggests that the naive subjects profited relatively little from location shift cues. All in all, performance remained quite poor, though perhaps somewhat better than expected.[5]

---

[5] This experiment was preceded by a pilot study of reversal discrimination, which was beset with "location shift" artifacts. However, the inexperienced subjects apparently did not profit much from this additional cue and performed poorly (59.1 percent correct), although somewhat better than predicted from the identification study in which these subjects had participated (53.8 percent correct). The most interesting result of the pilot study was the complete ineffectiveness of an additional independent variable: attenuation of one channel

TABLE 2: Monaural and dichotic reversal discrimination in Experiments IIA and IIB (percentages of correct responses). (Note: The underlines indicate probable artifacts due to "location shifts.")

	Stimuli	Monaural		Reversals	
		vs. 0	vs. 7	+0	+7
Experiment IIA	1	44.4	97.2	51.4	75.0
	2	66.7	100.0	55.6	73.6
	3	83.3	95.8	59.7	55.6
	4	91.7	94.6	62.5	55.6
	5	88.9	76.4	51.4	54.2
	6	91.7	48.6	72.2	58.3
	$\bar{X}$	77.8	85.4	58.8	62.1
Experiment IIB	1	57.5	93.7	61.2	62.5
	2	71.2	83.7	58.7	41.2
	3	86.2	91.2	57.5	57.5
	4	96.2	91.2	67.5	66.2
	5	90.0	63.7	60.0	48.7
	6	95.0	58.7	58.7	51.2
	$\bar{X}$	82.7	80.4	60.6	54.6

TABLE 3: Monaural controls and dichotic one-ear and crossover discrimination in Experiments IIA and IIB (percentages of correct responses).

	Stimuli	Monaural	One-ear		Crossover	
			+0	+7	+0	+7
Experiment IIA	1 vs. 3	76.4	63.9	59.7	68.7	72.2
	2 vs. 4	79.2	61.8	66.7	68.7	66.0
	3 vs. 5	79.2	59.0	68.1	66.7	60.4
	4 vs. 6	91.7	65.3	58.3	73.6	59.7
	$\bar{X}$	81.6	62.5	63.2	69.4	64.6
Experiment IIB	1 vs. 3	85.0	73.7	77.5	81.2	79.4
	2 vs. 4	78.7	71.2	68.7	69.4	66.9
	3 vs. 5	71.2	60.0	76.9	61.2	75.0
	4 vs. 6	87.5	73.7	76.2	76.2	84.4
	$\bar{X}$	80.6	69.7	74.8	70.0	76.4

by 10 dB (channel intensities at 85 and 75 dB). Although the fused syllables were lateralized toward the ear with the louder stimulus, the perceptual dominance of the louder syllable did not increase, and performance even decreased slightly. This is in agreement with the results of Cullen, Thompson, Hughes, Berlin, and Samson (1974) and Speaks and Bissonette (1975), who varied relative intensity in identification studies and obtained no effect in this range.

In Experiment IIB, performance in reversal discrimination correlated moderately (r = +0.45) with the absolute size of the ear advantage in one-ear discrimination, as predicted; however, the correlation did not reach significance. The variation in accuracy between different stimulus combinations followed no interpretable pattern.

One ear and crossover discrimination. These results are shown in Table 3. In Experiment IIB, 1 vs. 3 and 4 vs. 6, which were discriminated best monaurally, also showed the highest scores in one-ear and crossover discrimination, in agreement with Experiments IA and IB. (In Experiment IIA, there was no clear pattern.) Performance improved significantly ($p < .01$) from Experiment IIA to Experiment IIB. This suggests that location shifts played no role in these tasks in Experiment IIA, which agrees which subjective evidence and the comparison of Experiments IA and IB.

Crossover discrimination was slightly superior to one-ear discrimination ($p < .02$). The effect was more pronounced in Experiment IIA, but there was no significant interaction with experiments.

There was evidence of a REA in one-ear discrimination. Eight out of nine subjects in Experiment IIA and seven out of ten subjects in Experiment IIB showed a REA; one REA in Experiment IIA and two REAs and two LEAs in Experiment IIB were significant at the individual level. The average REAs corresponded to ø coefficients (Kuhn, 1973) of +0.07 ($p < .05$) in Experiment IIA and +0.05 ($p > .10$) in Experiment IIB. In the analysis of variance, the overall REA was only marginally significant ($p < .10$). However, there was a significant interaction with the nature of the constant stimulus ($p < .003$). As in Experiment IA, the REA was much larger when the constant stimulus was /bæ/ than when it was /gæ/. In fact, a small REA in Experiment IIA and a small LEA in Experiment IIB averaged out to zero in the +7 condition, while the +0 condition showed fairly large REAs in both experiments (ø = +0.10 and +0.15, respectively; both $p < .01$).

In crossover discrimination, there was also a marginally significant overall ear asymmetry ($p < .06$), which, however, occurred only in Experiment IIA: performance was higher when the acoustically more dissimilar stimuli were in the right ear. (For example, in 0+3 vs. 5+0, performance was higher when 0 and 5 were in the right ear.) This ear asymmetry in Experiment IIA corresponded to a ø coefficient of +0.09 ($p < .01$).

Discussion

The results showed reversal discrimination to be better than expected and crossover discrimination to be easier than one-ear discrimination. However, these effects were rather small and do not justify the conclusion that the listeners had access to the information in the separate channels prior to fusion. If one-ear discrimination were to be explained by the channel-accessibility hypothesis, reversal discrimination should have been considerably easier than one-ear discrimination. This was clearly not the case. The hypothesis also contradicts the subjective impression of perfect fusion and Repp's (1976b: Experiment IV) demonstration that selective attention to one ear is ineffective, and it must therefore be dismissed.

The small effects found were perhaps due to variations in ear dominance within subjects. It is quite conceivable that ear dominance is a rather unstable characteristic that exhibits fluctuations over time. Such variations around a mean value would aid reversal and crossover discrimination, especially in individuals with no strong ear asymmetries.

There was a REA in one-ear discrimination, but only when the constant stimulus was /bæ/, as in Experiment IA. This puzzling finding, together with the apparent unreliability of the REA and the tediousness of the task, does not make these discrimination tasks a promising alternative to dichotic identification tests as instruments for assessing ear advantages.

GENERAL DISCUSSION

What is the nature of the stimulus representations that the subjects tried to discriminate? They are not the phonetic labels of the dichotic fusions, because the obtained discrimination results did not conform to the predictions from the dichotic labeling functions (Experiments IA and IB). They are not the phonetic labels of the variable stimuli alone (prior to integration and fusion with the constant stimuli in the other channel), because accessibility of individual channels seems highly unlikely (Experiments IIA and IIB). Nor can they be "raw" auditory representations retained in some short-term store (Pisoni, 1971; Pisoni and Lazarus, 1974), since discrimination of low-level auditory codes would be expected to be more or less continuous and could not lead to the pronounced peaks and troughs in the one-ear discrimination functions (Experiments IA and IB). This leads to the conclusion that the codes that are discriminated must be an intermediate stage between initial auditory analysis and the final phonetic label, and that they most likely represent the integrated information from the two ears and not a single channel.

This intermediate stage can be more precisely specified within the framework of certain models of speech perception that postulate a limited number of discrete analyzing mechanisms that intervene between the auditory input and the phonetic label. These mechanisms may be termed "feature detectors" (Eimas and Corbit, 1973; Cooper, 1974; Cooper and Nager, 1975) or "prototypes" (Repp, 1976b); the distinction, while important in other contexts,[6] need not concern us here. Let us assume, then, that there are three "prototypes" corresponding to the three categories (B, D, G), and that each prototype exhibits maximal "sensitivity" to the acoustic input most appropriate for the corresponding categories. So, for example, a stimulus from the /bæ/ end of the place continuum will "activate" the B prototype most and the D and G prototypes only little; the next stimulus on the continuum, still heard as /bæ/ but closer to /dæ/ than the first syllable, will activate the B prototype a little less and the D prototype a little more, and so on. Such hypothetical activation values for the present stimuli (Table 1) are illustrated in Table 4 ("single-channel").[7]

---

[6] Repp, B. H. Dichotic competition of speech sounds: The role of acoustic stimulus structure. Unpublished manuscript.

[7] The degree of activation of a prototype most likely bears a nonlinear relationship to the acoustic "distance" between stimulus and prototype. The exact function will depend on the "response characteristic" of the prototype or on the "distribution of excitation" around the stimulus, about which little is known at the present time.

TABLE 4: Fictitious multicategorical vectors and one-step discriminability indices in single-channel and dichotic one-ear discrimination.

	Single-channel				+1				+7			
Stimuli	B	D	G	$(\sum d^2)^{1/2}$	B	D	G	$(\sum d^2)^{1/2}$	B	D	G	$(\sum d^2)^{1/2}$
1	8	1	1		8	1	1		4.5	1	4.5	
				1.4				0.7				0.7
2	7	2	1		7.5	1.5	1		4	1.5	4.5	
				6.5				2.9				2.9
3	2	6	2		5	3.5	1.5		1.5	3.5	5	
				1.4				0.7				0.7
4	1	7	2		4.5	4	1.5		1	4	5	
				4.2				2.1				2.1
5	1	4	5		4.5	2.5	3		1	2.5	6.5	
				2.8				1.0				1.0
6	1	2	7		4.5	1.5	4		1	1.5	7.5	
				1.4				0.7				0.7
7	1	1	8		4.5	1	4.5		1	1	8	

This representation of the stimulus information as a vector of prototype activation values has been termed "multicategorical" by Repp (1976b). The final category label is determined by a decision process that selects from the multicategorical vector the prototype with the highest activation level. We may assume that there is noise in the system, so that the decision process is probabilistic in nature. For the sake of simplicity, the numbers in Table 4 have been chosen to be roughly proportional to the probabilities of identification responses in the respective categories (according to the data in Repp, 1976b). They add up to a fixed sum for each stimulus, implying that each stimulus leads to the same degree of total activation in the system.

Let us now assume that a listener bases his discrimination responses not on the phonetic labels but on the multicategorical vectors, even in single-channel discrimination. An appropriate index for the discriminability of two vectors is the Euclidean distance between them (in a three-dimensional "prototype space," in the present example), which is equal to the square root of the sum of squared differences between corresponding elements. This discriminability index $(\sum d^2)^{1/2}$, is displayed for one-step discriminations in Table 4 ("single-channel"), as calculated from the hypothetical multicategorical vectors. It is evident that the index is maximal across category boundaries and minimal within categories, just like the obtained (and predicted) single-channel discrimination functions. Therefore, the assumption that listeners discriminate multicategorical vectors rather than phonetic labels is plausible and can, at least in principle, account for the categorical perception of single syllables.

We are now only one step removed from the explanation of the dichotic discrimination functions. In order to complete the argument, an assumption about the nature of dichotic interaction is necessary. Repp (1976b) has already argued from an analysis of the identification of dichotic fusions that dichotic integration of information takes place at the level of multicategorical representation and that the process is additive, that is, the multicategorical vector of a dichotic pair is the sum of the multicategorical vectors of the dichotic stimuli. When applied to our present problem, this leads immediately to the insight that the addition of a constant vector to each of two vectors does not change their discriminability, because it does not change the differences between corresponding elements and, hence, leaves the discriminability index unaffected. Therefore, one-ear discrimination functions should have the same

shape as single-channel discrimination functions, regardless of the nature of the constant stimulus. This was in fact obtained, at least in good approximation. However, the additivity assumption would predict no change in discriminability at all, whereas the obtained one-ear discrimination performance was considerably lower than single-channel performance. If we remember the assumption that the total amount of activation produced by a single syllable is constant and the fact that dichotic fusions sound like single syllables, it is then plausible that dichotic integration is not a simple summation but an averaging process that keeps the total activation constant. If each of two vectors is averaged with a constant vector, their relative discriminability will remain unchanged, but their absolute discriminability will decrease because averaging reduces all differences to half their original size. This is illustrated in Table 4 ("one-ear discrimination").

In order to account for ear dominance effects, we finally stipulate that dichotic integration consists in the weighted averaging of multicategorical vectors ($\underline{x}$, $\underline{y}$). The weights (a, b; $a+b = 1.0$) represent the relative dominance of each ear. Our model of dichotic integration is then: $a\underline{x} + b\underline{y} = z$.

This relatively simple model provides a good qualitative account of the data.[8] (A quantitative formulation is straightforward, and tests of a formal model are now in progress.) Note the dissociation of labeling and discrimination responses that occurs in dichotic fusions. By adding a constant stimulus to stimuli from a place continuum, the labeling functions are strongly biased toward the constant stimulus (cf. Table 4, assuming that the prototype activation values represent the probabilities of the corresponding responses; see also Repp, 1976b:Figure 1). On the other hand, discriminability remains independent of the constant stimulus and simply drops in absolute level, leaving the pattern unchanged. The fact that single-channel discrimination functions can be predicted from single-channel labeling functions may be a coincidence. The fact that even single-channel performance is usually somewhat better than predicted may be cited as additional (weak) evidence that discrimination is based not on the phonetic labels but on a lower-level representation.

---

[8] Those shifts in the discrimination peaks that were observed in Experiments IA and IB (primarily for the author as a subject) probably do not reflect individual tendencies to make some discriminations on the basis of phonetic labels, since it seems difficult to account for any peaks from phonetic discrimination alone (cf. the predicted dichotic functions in Figure 1, top). A finding from the earlier identification experiment is relevant here: the author showed a much stronger tendency toward "psychoacoustic fusions" (hearing D when /bæ/-/gæ/ is presented) than most other subjects. Repp (1976b) argued for an explanation of psychoacoustic fusions at the multicategorical level, but it may be that Cutting (1976) is right in hypothesizing a lower-level (probabilistic) auditory averaging process. Such auditory averaging, if it occurs, would precede the establishment of the (single) multicategorical code, and it would destroy additivity and result in a shift of discrimination peaks. The finding that primarily the author showed such shifts and that they occurred especially in the region of /bæ/-/gæ/ contrasts (cf. also Figure 1) supports this explanation. Therefore, in order to account for the detailed response pattern, a two-stage model may be necessary. It seems, however, that auditory averaging plays only a minor role for most subjects.

The present model makes the distinction between phonetic (discrete) and auditory (continuous) discrimination unnecessary, at least in the present context (Fujisaki and Kawashima, 1970; Pisoni, 1971; Pisoni and Lazarus, 1974). The multicategorical vector is a code consisting of several discrete elements that assume continuous values, and it is therefore both discrete and continuous. More sensitive discrimination tasks will lead to better performance than less sensitive ones (Pisoni and Lazarus, 1974) by changing the criterion for the detection of differences between vectors; it is no longer necessary to invoke auditory memory to account for this finding. Most likely, the multicategorical vector is also the basis for confidence judgments and ratings of category goodness (Barclay, 1972; Vinegrad, 1972; Summerfield, 1975; Cooper, Ebert, and Cole, 1976). It is useful to consider the multicategorical vector as the stimulus code on which the human listener operates according to the demands of the task. Deciding upon phonetic labels is only one of these possible tasks, and other tasks such as discrimination or rating are probably not more based on labels than identification is based on implicit discriminations or ratings. The notion of an intermediate, "multicategorical" stage may contribute to the understanding of various problems in speech perception that so far have been viewed in the light of the ubiquitous auditory-phonetic dichotomy (Studdert-Kennedy, in press).

REFERENCES

Barclay, J. R. (1972) Noncategorical perception of a voiced stop: A replication. Percept. Psychophys. 11, 269-273.

Cherry, E. C. and B. McA. Sayers. (1956) "Human 'cross-correlator'": A technique for measuring certain parameters of speech perception. J. Acoust. Soc. Am. 28, 889-895.

Cooper, W. E. (1974) Adaptation of phonetic feature analyzers for place of articulation. J. Acoust. Soc. Am. 56, 617-627.

Cooper, W. E., R. R. Ebert, and R. A. Cole. (1976) Perceptual analysis of stop consonants and glides. J. Exp. Psychol.: Human Perception and Performance 2, 92-104.

Cooper, W. E. and R. N. Nager. (1975) Perceptuo-motor adaptation to speech: An analysis of bisyllabic utterances and a neural model. J. Acoust. Soc. Am. 58, 256-265.

Cullen, J. K., Jr., C. L. Thompson, L. F. Hughes, C. I. Berlin, and D. S. Samson. (1974) The effects of varied acoustic parameters on performance in dichotic speech perception tasks. Brain Lang. 1, 307-322.

Cutting, J. E. (1976) Auditory and linguistic processes in speech perception: Inferences from six fusions in dichotic listening. Psych. Rev. 83, 114-140.

Eimas, P. D. (1963) The relation between identification and discrimination along speech and non-speech continua. Lang. Speech 6, 206-217.

Eimas, P. D. and J. D. Corbit. (1973) Selective adaptation to linguistic feature detectors. Cog. Psychol. 4, 99-109.

Fujisaki, H. and T. Kawashima. (1970) Some experiments on speech perception and a model for the perceptual mechanism. In Annual Report of the Engineering Research Institute (Faculty of Engineering, University of Tokyo) 29, 207-214.

Halwes, T. G. (1969) Effects of dichotic fusion on the perception of speech. Unpublished Ph.D. dissertation, University of Minnesota.

Kuhn, G. M. (1973) The Phi coefficient as an index of ear differences in dichotic listening. Cortex 9, 447-457.

Liberman, A. M., K. S. Harris, H. S. Hoffman, and B. C. Griffith. (1957) The discrimination of speech sounds within and across phoneme boundaries. J. Exp. Psychol. 54, 358-368.

Pisoni, D. B. (1971) On the nature of categorical perception of speech sounds. Unpublished Ph.D. dissertation, University of Michigan.

Pisoni, D. B. and J. H. Lazarus. (1974) Categorical and noncategorical modes of speech perception along the voicing continuum. J. Acoust. Soc. Am. 55, 328-333.

Pollack, I. and D. B. Pisoni. (1971) On the comparison between identification and discrimination tests in speech perception. Psychon. Sci. 24, 299-300.

Repp, B. H. (1976a) Effects of fundamental frequency contrast on discrimination and identification of dichotic CV syllables at various temporal delays. Mem. Cog. 4, 75-90.

Repp, B. H. (1976b) Identification of dichotic fusions. Haskins Laboratories Status Report on Speech Research SR-45/46 (this issue). [Also J. Acoust. Soc. Am. (in press).]

Speaks, C. and L. J. Bissonette. (1975) Interaural-intensive differences and dichotic listening. J. Acoust. Soc. Am. 58, 893-898.

Studdert-Kennedy, M. (in press) Speech perception. In Contemporary Issues in Experimental Phonetics, ed. by N. J. Lass. (New York: Academic Press).

Summerfield, A. Q. (1975) Cues, contexts, and complications in the perception of voicing contrasts. Speech Perception, Report on Speech Research in Progress (Psychology Department, The Queen's University of Belfast) Series 2, no. 4, 99-130.

Vinegrad, M. D. (1972) A direct magnitude scaling method to investigate categorical versus continuous modes of speech perception. Lang. Speech 15, 114-121.

Young, L. L., Jr., C. Parker, and R. Carhart. (1975) Coherence function for speech. J. Acoust. Soc. Am., Suppl. 58, S54(A).

Coperception: Two Further Preliminary Studies

Bruno H. Repp*

ABSTRACT

Two "same-different" reaction-time studies were conducted to investigate the temporal limits of perceptual integration ("coperception") in speech perception, as measured by the influence of irrelevant context on the latencies of judgments about designated "target" segments. The first study varied the duration of the silent closure period of a medial stop in synthetic vowel-consonant-vowel (VCV) syllables. Surprisingly, the implosive and explosive transitions of the stop consonant (the target) were "coperceived," together with the final vowel (the irrelevant context), over as much as 200 msec of intervening silence. The second study varied the duration of the vowel in VC syllables and found no coperception of the vowel (target) with the final consonant (context) at all. Both results may reflect the degree of discriminability of the particular target phonemes and contexts used, so that further studies will be necessary to determine the generality of the present findings and to elucidate the role of discriminability in coperception.

INTRODUCTION

In an earlier report (Repp, 1975), I introduced the term "coperception" to denote a certain class of contextual effects in speech perception. Coperception was defined, in analogy to coarticulation, as the influence of one (phonemic) segment on the perception of another (phonemic) segment in an utterance. The measure of perception was stipulated to be reaction time (of same-different judgments, classification, or detection); that is, it was presupposed that the speech signal is fully intelligible. This excludes phenomena such as masking from the definition of coperception.

The notion of coperception is a direct extension of Garner's (1974) concept of stimulus integrality to the temporal domain. An extension to the spatial domain in vision has recently been undertaken by Pomerantz and Garner (1973) and Pomerantz and Schwaitzberg (1975) whose work provides a parallel to the present

---

*Also University of Connecticut Health Center, Farmington.

Acknowledgment: This research would not have been possible without the generous hospitality of Haskins Laboratories and its director, Alvin Liberman. The author was supported by NIH Grant T22 DE00202 to the University of Connecticut Health Center.

[HASKINS LABORATORIES: Status Report on Speech Research SR-45/46 (1976)]

approach. Garner's theory and methods have been applied extensively to the perception of the simultaneously present dimensions of single stimuli (cf. Garner and Felfoldy, 1970, and Garner, 1974, in vision; Wood, 1975a, 1975b, in speech perception). The problem may be formulated in terms of selective attention (e.g., Wood and Day, 1975): a stimulus is called integral if its individual dimensions or components cannot be attended to without taking its other dimensions into account. A speech signal is a multidimensional auditory event that extends over time, just as visual stimuli extend into space. In both cases, there are obvious limits to stimulus integrality, or coperception. When two visual stimuli are sufficiently separated in space, they will cease to be integral (Pomerantz and Schwaitzberg, 1975). Likewise, if two speech segments are sufficiently separated in time, they will no longer be coperceived. By varying the spatial or temporal structure of the stimulus events, the factors that lead to coperception in the appropriate modalities may be explored. In vision, there is good reason to believe that our intuitions about what forms a good Gestalt will be relevant (Pomerantz and Schwaitzberg, 1975). We may ask the analogous question in speech perception, and in auditory perception in general: What portions of the auditory signal represent a "Gestalt," and what are the properties that define it?

Pisoni and Tash (1974) and Wood and Day (1975) have demonstrated that an initial stop consonant and the following vowel are such an auditory Gestalt. The relevant factor here may be the absence of any acoustic segmentation corresponding to the two phonetic segments, especially when the initial stop consonant is voiced and has no "burst," that is, when it is represented only by the initial transitions of the vowel. In other words, the continuity of the signal may be a crucial factor in coperception, as it is in the perception of temporal order (Dorman, Cutting, and Raphael, 1975). However, consider a medial stop consonant, as in /abi/. Here, the implosive transitions of the initial vowel are separated from the explosive transition into the final vowel by a silent closure period. (In natural speech, low-intensity voicing may continue through the closure.) Are the two portions of the auditory signal, which separately are heard as /ab/ and /bi/, still coperceived across the gap separating them? It has been demonstrated (Repp, 1975) that they are, as is intuitively suggested by the fact that only a single consonant is heard.

The present paper reports two further preliminary studies. They are considered preliminary because their results suggest additional factors that will have to be taken into account in research on coperception. Therefore, these studies will primarily serve to illustrate and discuss some methodological issues. Their results cannot be considered conclusive.

The first experiment was concerned with the limits of the coperception effect in vowel-consonant-vowel (VCV) syllables: If the silent closure period is extended in duration, when will coperception of implosive and explosive transitions (plus the final vowel) cease? The prediction was straightforward: at a certain separation, not one but two (geminate) consonants will be heard, for example, /ab-bi/ (Delattre, 1971; Dorman, Raphael, Liberman, and Repp, 1975), and this closure duration was expected to mark the end of coperception.

The second study investigated coperception in vowel-consonant (VC) syllables. Pisoni and Tash (1974) and Wood and Day (1975) have shown that, in consonant-vowel (CV) syllables, judgments about the vowel are influenced by variations in the initial consonant, although the vowel has a steady state that is entirely

independent of the consonant. Apparently, the fact that the consonant (i.e., the formant transitions) precedes the vowel is important here. In VC syllables, on the other hand, the steady state of the vowel precedes the final consonant. Will the consonant still be coperceived with the vowel? Clearly, if the vowel is sufficiently long, a response to the vowel can be made before the final transitions even enter the ear, so there must be a limit to coperception. In order to investigate this limit, the duration of the vowel was varied systematically.

Both studies reported here used same-different paradigms, based on the assumption that the results would be comparable to those obtained in a speeded-classification paradigm, the more traditional technique for assessing stimulus integrality (Garner, 1974). While there is little evidence to suggest the contrary, the two paradigms nevertheless differ in important respects, and it will be necessary to compare the two techniques in future studies. In the same-different task, two utterances are presented in succession, and the listener is asked to judge whether a certain well-defined segment is the same or different in the two stimuli, while other irrelevant segments vary randomly. Coperception is said to exist when "same" judgments are facilitated by identity of the contexts (relative to nonidentical contexts) and/or when "different" judgments are facilitated by nonidentity of the contexts (relative to identical contexts).

EXPERIMENT I

Method

Subjects. Eight paid volunteers participated. All were native speakers of English, had normal hearing and little experience in reaction-time tasks.

Stimuli. Four VCV syllables--/abi/, /adi/, /abɛ/, and /adɛ/--were synthesized on the Haskins Laboratories parallel resonance synthesizer. They consisted of two acoustic segments, 200 and 300 msec long, respectively, separated by a variable silent gap. The first segment included an initial steady state followed by 45-msec implosive transitions that did not vary with the final vowel (i.e., they were identical in /abi/ and /abɛ/ and in /adi/ and /adɛ/). The second segment began with 45-msec explosive transitions (independent of the initial vowel) and ended in a steady state. The durations of the silent closure period were 50, 100, 150, and 200 msec, resulting in total stimulus durations of 550, 600, 650, and 700 msec, respectively.

An experimental tape containing pairs of these stimuli was recorded using the pulse code modulation system at Haskins Laboratories. The stimulus onset asynchrony within a pair was 1 sec and constant (the interstimulus interval varied with the duration of the closure period), and the interpair interval was 3 sec. The tape contained first a short practice series (eight pairs at each closure duration), which was followed by four blocks of 80 pairs each. Each block corresponded to a particular closure duration and contained five subblocks (not separated by pauses), each containing the 16 possible combinations of the four syllables in random order.

Procedure. The subjects were tested individually in a single session lasting about 90 minutes. Each subject listened to the experimental tape twice. The four blocks were presented once in ascending order (i.e., with closure duration increasing) and once in descending order, counterbalanced between subjects.

The assignment of the hands to the two response keys ("same-different") was also counterbalanced between subjects. The subjects were instructed to respond as quickly and as accurately as possible. Before the experiment, they were told exactly what the stimuli represented, that they would tend to hear two identical consonants at the longest closure duration(s), and that they should judge the consonants only and ignore the variation in the final vowel. It was stressed that they should respond as soon as they could reach a decision and not wait for the end of the utterance.

The tape was played back from an Ampex AG-500 tape recorder through a mixer to Telephonics TDH-39 earphones. The intensity was set at a comfortable level (about 75 dB SPL). The syllables, which had been recorded on separate channels, were presented binaurally after electronically mixing the two channels. The onset of the first syllable in a pair triggered a Hewlett-Packard 522B electronic counter that was stopped by the subject's depression of one of the two response keys. The reaction time was recorded to the nearest millisecond, together with the kind of response given.

The stimulus onset asynchrony (1 sec) was subtracted from the reaction times, so that they were measured from the onset of the second syllable in a pair. (This is how the reaction times are given below. In order to obtain the latencies with reference to the onset of the silent closure period in the second syllable--as in Repp, 1975--another 200 msec should be subtracted.) Prior to analysis, median reaction times were calculated for the five replications of the same stimulus pair in each block, omitting errors. Further analysis was in terms of the means of these medians.

Results and Discussion

Assuming that the basic effect of coperception is replicated at the shortest closure duration (50 msec), there are two patterns the results may follow. If the subjects were able to rely increasingly on the implosive transitions alone as closure duration was lengthened, the difference between reaction times as a function of context should decrease to zero, and the absolute latencies should not be affected by closure duration. This was the expected outcome. On the other hand, the null hypothesis is that at all closure intervals the subjects would rely on the explosive transitions alone. In this case, not only should the context effect remain constant, but the latencies should increase as a linear function of closure duration with a slope of unity. This is because latencies are measured from the onset of the VCV syllable, and an increase in closure duration means that the listener has to wait that much longer before he hears the explosive transitions and can reach a decision.

The outcome is shown in Figure 1. Surprisingly, it is in close agreement with the null hypothesis. It can be seen that all reaction times increased with slopes close to unity, especially at the longer closure durations; the flatter slopes at the short durations probably reflect a floor effect. It is also evident that at all closure durations "same" reaction times were faster when the final vowel was the same than when it was different; that is, coperception was present and persisted up to the longest interval. Only the "different" reaction times show an interaction: at the shortest closure duration, they were faster when the final vowels were different, as predicted; but there was no such difference at the longer durations. "Different" reaction times were considerably slower than "same" reaction times, which is a common finding in tasks of this sort.

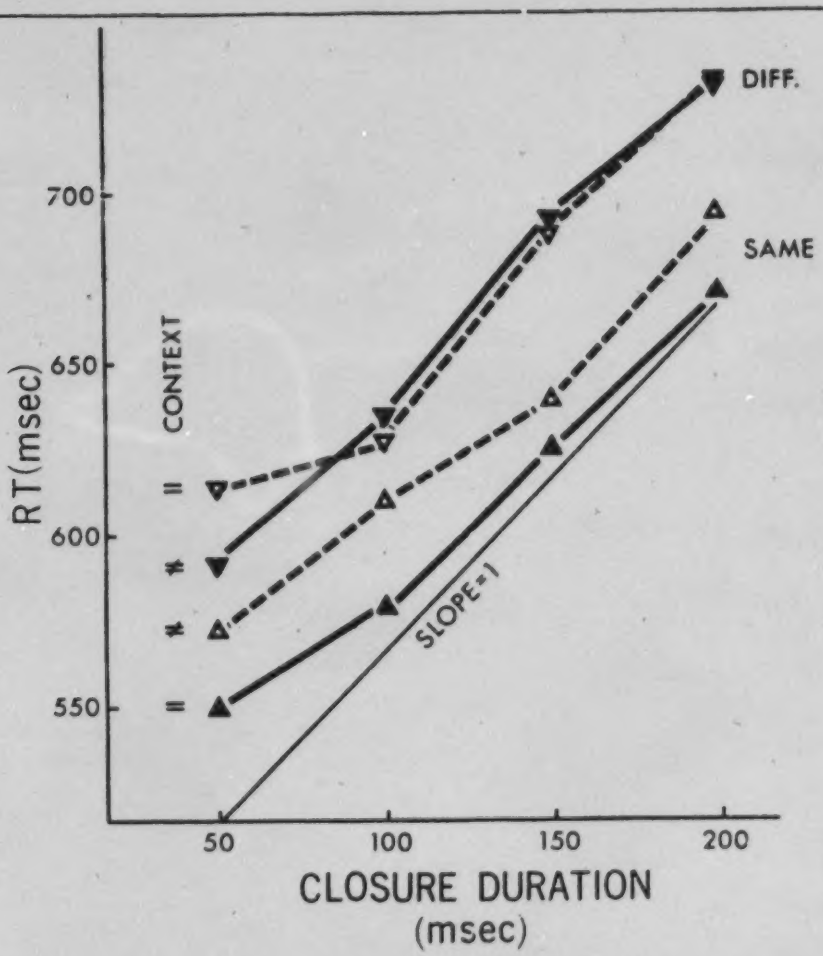


Figure 1: Average median reaction times as a function of closure duration. "Same" and "different" latencies are shown for identical and non-identical vowel contexts.

Figure 1 is actually not very representative of the individual data, which showed substantial variation. In view of this variation, and of the negative result, no statistical analysis was deemed necessary. The data were considerably more variable than in the previous similar experiment (Repp, 1975). The only effect consistently shown by all subjects was the linear increase in reaction times with closure duration. Only four of the eight subjects actually showed a positive coperception effect in their "same" reaction times (but then a very large one, which accounts for the average positive effect). This is in contrast to the previous results (Repp, 1975) where all 12 subjects showed a positive effect. The coperception effect on "different" reaction times was similarly variable, and there were also surprisingly large variations within the data of individual subjects. Consequently, the only reliable finding exhibited in Figure 1 is the linear increase in reaction times. However, this result is sufficient to suggest that all listeners made their judgments on the basis of the explosive transitions alone.

It is interesting to observe that the error pattern did not closely correlate with the latencies. The average error rate was 6.3 percent, with individuals varying between 1.8 and 12.0 percent. The errors decreased by about one-third from the first to the second half of the session. The pattern is shown in Table 1.

TABLE 1: Average error percentages as a function of closure duration and type of stimulus pair.

Correct response	Context	Closure duration 50	100	150	200	$\bar{X}$
"Same"	=	3.4	3.1	1.9	5.3	3.4
	≠	10.3	10.0	9.1	7.8	9.3
"Different"	=	8.8	5.0	10.0	5.0	7.2
	≠	5.3	3.1	7.2	5.0	5.2
	$\bar{X}$	7.0	5.3	7.1	5.8	6.3

It can be seen that the error rates at closure durations of 100 and 200 msec were lower than at 50 and 150 msec; by no means did the errors follow the linear increase observed for the reaction times. The fluctuation was due to incorrect "same" responses (lower part of Table 1); the frequencies of incorrect "different" responses were fairly constant. The effect of context is reflected in the error frequencies: there were fewer incorrect "different" responses when the context was the same than when it was different, and fewer incorrect "same" responses when the context was different than when it was the same. The first effect was present at all closure durations but reduced at 200 msec, while the second effect was less pronounced but present at all closure durations except 200 msec. Thus, the error rates are suggestive of some change in processing at the longest closure duration.

It would be naive to take the results at face value and conclude that implosive and explosive transitions are perceptually integrated over a total period of almost 300 msec, even if this is still within the upper limits of the acoustic store postulated by Massaro (1972, 1974). It is also unlikely that all eight subjects failed to obey the instructions, which were clear enough. One possibility is that the prediction that coperception would cease as soon as geminate consonants are heard was essentially correct but that the naive subjects still heard only a single consonant at the longest closure duration. The closure durations were selected by the author, who clearly heard geminate consonants with the 200-msec closure period but not at the shorter durations. It is a shortcoming of the experiment that "single versus geminate" judgments were not elicited in a control condition.

Another deficiency of this study was that it did not test whether the subjects actually were able to discriminate the implosive transitions in isolation. Clearly, if they could not tell /ab/ and /ad/ apart, they would have had to rely on the explosive transitions in the VCVs. However, four of the subjects participated in Experiment II, described below, which included VC syllables identical

with the first portion of the VCV syllables of the present study. Three subjects were able to discriminate them without much difficulty, and only one subject failed. Since all subjects showed a linear increase in reaction time in the present experiment, failures to discriminate the implosive transitions are not a likely explanation for the results.

However, it is well-known to those working with synthetic speech that implosive transitions are not easy to discriminate, especially in the absence of the release burst that natural syllable-final stops often show. This low "salience" of implosive transitions may be the reason that the explosive transitions determine the perceived place of articulation when implosive and explosive transitions in VCV syllables are artificially brought into conflict (Dorman, Raphael, Liberman, and Repp, 1975; Fujimura, 1975). It also may lie at the heart of the present problem. Difficult discriminations are necessarily associated with long decision times. Let us assume that the subjects heard the implosive transitions at the longest closure duration; that is, they heard geminate consonants. It is then possible that they attempted to reach a decision as soon as they heard the syllable-final stop, but that the decision process was not yet completed by the time the syllable-initial stop arrived. This may have interrupted the ongoing decision process, or it may have initiated a separate desision process of its own, which overtook the earlier process. For example, if the decisions for the syllable-final consonants lasted about 300 msec longer than the decisions for the syllable-initial consonants, on the average, the latter would have been completed earlier than the former in most cases. Although such a large difference is rather unlikely, the hypothesis needs to be tested by asking subjects to discriminate implosive transitions in isolation (i.e., in VC syllables).[1]

The tentative conclusion from the preceding paragraph is that the discriminability of the target segments may play an important role and should be included as a parameter in studies of coperception, whenever possible. Wood and Day (1975) have discussed the same problem in the context of the speeded-classification paradigm. Unfortunately, in the case of syllable-final transitions, not much can be done to improve discriminability. Perhaps the /ab/-/ag/ contrast will prove easier to discriminate than the /ab/-/ad/ contrast, because of the larger acoustic difference in the transitions. In addition, a future experiment might employ VCV and VC syllables in the same design, which should direct the subjects' attention to the syllable-final consonants. Practice in discriminating implosive transitions may also reduce the difficulty of the task. Finally, the explosive transitions could be made less discriminable by making them acoustically more similar, in order to increase the corresponding decision times and

---

[1] In fact, this suggests an alternative explanation of the increase in reaction times with closure period duration. It may be that the subjects did make decisions on the basis of the implosive transitions with a certain probability that increased with closure duration (perhaps only at the longest closure duration). The reaction times would then represent a mixture of two distributions--slow latencies for implosive transitions and fast latencies for explosive transitions--and the increase with closure duration would represent an increase in the proportion of slow latencies. However, it would be a rare coincidence if this kind of process had produced the linear functions shown in Figure 1, and one should also have expected an increase in errors and a decrease in the coperception effect as closure period was lengthened. Therefore, the explanation seems rather unlikely.

to discourage the subjects from relying too much on the syllable-initial consonants (if subjective strategies are involved at all). These approaches will have to be tried out in future experiments.

EXPERIMENT II

Method

Subjects. There were four subjects who had previously participated in Experiment I.

Stimuli. The stimuli were the syllables /ab/, /ad/, /ɛb/, and /ɛd/, synthesized on the Haskins Laboratories parallel resonance synthesizer. The final transitions (45 msec) were preceded by a steady-state vowel of variable duration. The total syllable durations were 100, 150, 200, 250, and 300 msec.

The experimental tape first contained a brief practice list of single syllables for identification. It was followed by five blocks of 160 pairs each. The 160 pairs consisted of the 16 possible combinations of the four syllables, with two possible durations of the first syllable (150 or 250 msec) and five possible durations of the second syllable, which were completely randomized. The stimulus onset asynchrony was constant at 750 msec, and the interpair interval was 3 sec.

Procedure. The procedure was similar to that in Experiment I, except that the subjects were instructed to judge as rapidly as possible whether the vowels were the same or different, ignoring vowel duration and the final consonant. Reaction times were measured from the onset of the second syllable in a pair. The analysis was performed on mean reaction times, omitting errors and exceptionally long latencies.

Results and Discussion

Two hypotheses were tested in this experiment. One predicted that there would be a coperception effect when the second syllable in a pair was sufficiently short, and that this effect would disappear as the duration of the second syllable was increased. The second hypothesis predicted, on the assumption that fairly literal representations of the speech sounds are compared in the brain, that reaction times (perhaps "same" responses only) would be shorter when the first and the second syllable had the same duration. The results are shown in Table 2.

Table 2 shows that the reaction times exhibited surprisingly little variation (which attests to the reliability of the data). Neither hypothesis was supported. There was no indication of any coperception effect, nor was there any interaction with syllable duration. The only consistent difference was between "same" and "different" latencies, a trivial finding. Although the results were based on only four subjects, it seemed useless to run further subjects in this task.

The average error rate was 4.9 percent. The pattern of errors with respect to syllable duration is shown in Table 3.

TABLE 2: Mean reaction time as a function of syllable durations and type of stimulus pair.

		Syllable duration									
	First:	150					250				
	Second:	100	150	200	250	300	100	150	200	250	300
Response	Context										
"Same"	=	337	333	350	343	353	335	352	363	345	327
	≠	340	335	338	334	356	358	335	317	323	333
"Different"	=	366	356	340	366	343	357	349	374	357	365
	≠	363	353	373	359	361	354	350	358	351	387

TABLE 3: Average error percentages as a function of syllable durations.

		Duration of second syllable					
		100	150	200	250	300	$\bar{X}$
Duration of first syllable	150	7.2	4.7	3.4	2.2	2.2	3.9
	250	9.7	6.3	5.9	4.1	3.1	5.8
	$\bar{X}$	8.4	5.5	4.7	3.1	2.7	4.9

In contrast to the latencies, the error rates declined steadily as the duration of the second syllable increased, but, surprisingly, they were higher with the longer duration of the first syllable. At the shortest duration of the second syllable, the error pattern was in agreement with a coperception effect (not shown in Table 3), but there were too few observations to draw any conclusions (and, moreover, coperception is defined in terms of reaction times, although the error frequencies often show a positive correlation with the latencies). No statistical analysis was conducted.

Why did the reaction times show no effect? The reason may be that only the vowel onset matters and the information that follows is irrelevant. In other words, final consonants may not be coperceived with the preceding vowel. Such a conclusion would be highly interesting, but the present data do not justify it yet. Rather, it is likely that discriminability again played a role. The vowel discrimination was fairly easy, so that the decisions may have been completed before the final consonant was processed. In addition, the final transitions were not easy to discriminate, so that they were processed more slowly and therefore could not affect the vowel decision any more. In order to have any detectable effect, the context must be highly discriminable. It is planned to repeat the experiment with VCV syllables and more similar (initial) vowel targets. This should both increase the decision times for the vowel targets and decrease the decision times for the following consonants (medial consonants are probably more

discriminable than final stops), which should improve the sensitivity of the experiment.

A recent study by Healy and Cutting (in press) illustrates the problem of target discriminability. They used a detection paradigm in which a subject hears a list of utterances and responds to only one of them. They presented isolated vowels and VC syllables and asked the subjects to detect either a vowel or a VC syllable. Their subsequent comparison of vowel and syllable detection latencies showed faster syllable detection latencies for vowels that were difficult to discriminate (in a control condition) but faster vowel detection latencies for vowels that were easy to discriminate. This provides evidence that the final consonant may be coperceived with the preceding vowel, given that the vowel is difficult to classify. Suggestive evidence comes also from a recent study by Strange, Jenkins, and Edman (1975), who found that the intelligibility of isolated vowels increases when they are followed by a stop consonant, although, in this case, perceptual integration may have occurred at a later stage. It is likely that a more sensitive experiment than the present one will show coperception in VC syllables.

CONCLUSIONS

While the results of the present experiments are not conclusive, they have been helpful in pointing out a methodological issue, perhaps more so than "positive" outcome. Nor are the results invalid; they merely represent a sample from a whole continuum of stimulus discriminability. The discriminability of both the target and the context will have to be a parameter in future studies of coperception. It is likely that the limits of temporal integration in speech perception depend on the ease of discrimination of successive portions of the speech signal. If this is true, it means that there are no fixed "units" that are processed successively but that a number of concurrent and overlapping processes are triggered by the acoustic stimulus. The size of these processing units depends on the clarity of the information. In other words, the speech processor "accumulates evidence" until it can reach a decision. However, while this may accurately describe its operation in reaction-time tasks, generalizations to the processing of natural speech must be made with caution, because the target of attention is usually not at the phonemic level. Coperception studies reveal only the lower limits of perceptual integration, not its upper limits, which may be at least as important in "normal" speech perception.

REFERENCES

Delattre, P. (1971) Consonant gemination in four languages: An acoustic, perceptual and radiographic study. International Review of Applied Linguistics 9, 31-52; 9, 97-113.

Dorman, M. F., J. E. Cutting, and L. J. Raphael. (1975) Perception of temporal order in vowel sequences with and without formant transitions. J. Exp. Psychol.: Human Perception and Performance 1, 121-129.

Dorman, M. F., L. J. Raphael, A. M. Liberman, and B. H. Repp. (1975) Some maskinglike phenomena in speech perception. Haskins Laboratories Status Report on Speech Research SR-42/43, 265-276.

Fujimura, O. (1975) A look into the effects of context: Some articulatory and perceptual findings. Paper presented at the 8th International Congress of Phonetic Science, Leeds, England, 17-23 August.

Garner, W. R. (1974) The Processing of Information and Structure. (Potomac, Md.: Lawrence Erlbaum Assoc.).

Garner, W. R. and G. L. Felfoldy. (1970) Integrality of stimulus dimensions in various types of information processing. Cog. Psychol. 1, 225-241.

Healy, A. F. and J. E. Cutting. (in press) Units of speech perception: Phoneme and syllable. J. Verbal Learn. Verbal Behav.

Massaro, D. W. (1972) Preperceptual images, processing time, and perceptual units in auditory perception. Psychol. Rev. 79, 124-145.

Massaro, D. W. (1974) Perceptual units in speech recognition. J. Exp. Psychol. 102, 199-208.

Pisoni, D. B. and J. Tash. (1974) "Same-different" reaction times to consonants, vowels, and syllables. In Research on Speech Perception (Department of Psychology, Indiana University), Progress Report No. 1.

Pomerantz, J. R. and W. R. Garner. (1973) Stimulus configuration in selective attention tasks. Percept. Psychophys. 14, 565-569.

Pomerantz, J. R. and S. D. Schwaitzberg. (1975) Grouping by proximity: Selective attention measures. Percept. Psychophys. 18, 355-361.

Repp, B. H. (1975) "Coperception": A preliminary study. Haskins Laboratories Status Report on Speech Research SR-42/43, 147-157.

Strange, W., J. J. Jenkins, and T. Edman. (1975) Identification of vowels in CV and VC syllables. J. Acoust. Soc. Am., Suppl. 58, S59(A).

Wood, C. C. (1975a) Auditory and phonetic levels of processing in speech perception. J. Exp. Psychol.: Human Perception and Performance 1, 3-20.

Wood, C. C. (1975b) A normative model for redundancy gains in speech discrimination. In Cognitive Theory: Volume 1, ed. by F. Restle, R. M. Shiffrin, N. J. Castellan, H. Lindman, and D. B. Pisoni. (Potomac, Md.: Lawrence Erlbaum Assoc.).

Wood, C. C. and R. S. Day. (1975) Failure of selective attention to phonetic segments in consonant-vowel syllables. Percept. Psychophys. 17, 346-350.

"Posner's Paradigm" and Categorical Perception: A Negative Study

Bruno H. Repp*

ABSTRACT

A reaction-time study was conducted with four synthetic syllables from a "place continuum" (/bae/-/dae/$_1$-/dae/$_2$-/gae/). A special counterbalanced design was used to assess the effect of acoustic similarity on reaction time. The study included a "same-different" and a classification task, two different temporal delays between the syllables, and binaural versus dichotic (i.e., alternating monaural) presentation. However, no effects of auditory similarity were found, which contradicts a recent study by Eimas and Miller (1975) that used similar stimuli.

INTRODUCTION

Posner and Mitchell (1967) introduced an experimental paradigm that has led to some of the most elegant and successful research in visual information processing (e.g., Posner, 1969; Posner, Boies, Eichelman, and Taylor, 1969). The task consists in judging whether two letters are the same or different, with reaction time as the dependent variable. The two letters can be either identical (AA) or different (AB); in addition, they can have the same name but be physically different (Aa). The subjects are instructed to respond "same" when the two letters have the same name, and "different" otherwise. The principal finding is that "same" reaction times are faster for pairs that are physically identical (AA) than for pairs that are physically different (Aa). This suggests that physically identical letters can be matched at an earlier "node" in processing, which uses purely visual information, while name matches in the absence of physical identity take place at a later processing stage. Posner and Keele (1967) introduced temporal delays between the two stimuli, in order to find out whether the visual information that leads to the relative advantage for physical matches is subject to decay. They found a steady decline of the reaction-time difference over the first 2 sec, suggesting that the visual information is held in a relatively short-lived store.

Similar paradigms have been profitably applied in speech perception (e.g., Springer, 1973; Cole, Coltheart, and Allard, 1974; Repp, 1976a). Perhaps the

---

*Also University of Connecticut Health Center, Farmington.

Acknowledgment: This research was made possible by the generous hospitality of Haskins Laboratories and its director, Alvin Liberman. The author was supported by NIH Grant T22 DE00202 to the University of Connecticut Health Center.

[HASKINS LABORATORIES: Status Report on Speech Research SR-45/46 (1976)]

most interesting of these studies is that of Pisoni and Tash (1974). They applied Posner's paradigm to the classical problem of categorical perception. It is well-known that initial stop consonants are easy to discriminate as long as they are perceived as belonging to different categories, but that acoustic differences within these categories are almost impossible to detect (e.g., Pisoni, 1971). It has been suggested that this phenomenon may be due to the rapid loss of auditory information from memory (Fujisaki and Kawashima, 1970; Pisoni, 1971, 1973). Pisoni and Tash (1974) presented two synthetic syllables in close succession, which could be either physically identical (e.g., /ba/$_1$-/ba/$_1$), different acoustically but belonging to the same category (/ba/$_1$-/ba/$_2$), or belonging to different categories (/ba/-/pa/). The acoustic variable was voice onset time (VOT), the most important cue for the distinction between /ba/ and /pa/. The listeners were not aware of these acoustic variations and simply made "same-different" judgments with respect to the categories of the syllables. Pisoni and Tash found significantly shorter "same" reaction times for "physical matches" than for mere "name matches," just as Posner did. In addition, they found "different" reaction times to decrease with the acoustic difference between two syllables from different categories, which constitutes additional evidence for the availability of auditory information. (The corresponding finding in vision would be faster "different" latencies for Ab pairs than for AB pairs, a condition that has rarely been included and then has not yielded a positive effect--e.g., Besner and Coltheart, 1975). Pisoni and Tash suggested a two-stage processing model that allows for fast auditory matches to be conducted before slower phonetic (name) matches. Stimuli that are either identical or very different from each other may permit lower-level auditory decisions, while more ambiguous cases are decided at the phonetic level.

The Pisoni and Tash findings are especially interesting because, in contrast to other Posner-type tasks, the subjects are not aware of the physical differences within name categories; that is, no special "name match" instructions are necessary, as in the letter-matching task. Again, the question arises whether and how fast the auditory information is lost from memory. This may be investigated by varying the interval between the two syllables that are to be compared. I conducted such a study two years ago at the University of Chicago.[1] Pairs of syllables from a VOT continuum (ranging from /ba/ to /pa/, as in Pisoni and Tash, 1974) were presented at stimulus onset asynchronies (SOAs) between 0 and 3.3 sec. There was a clear effect of acoustic differences on "different" reaction times, which, moreover, did not decrease as the delay between the syllables increased. However, in contrast to the findings of Pisoni and Tash (1974), there was no clear evidence of any effect on "same" reaction times, which is the primary evidence for the availability of auditory information.

The effect on "different" reaction times could have a different explanation. It is well-known that it takes longer to classify stimuli that lie close to a category boundary than stimuli that are far from the boundary (Studdert-Kennedy, Liberman, and Stevens, 1963; Pisoni and Tash, 1974; Eimas and Miller, 1975; Repp, 1975). Pairs of acoustically very discrepant stimuli necessarily contain stimuli from the ends of the acoustic continuum, while pairs of acoustically

---

[1] Repp, B. H. (1974) Categorical perception, auditory memory, and dichotic interference. Unpublished manuscript. Copies of this paper are available from the author upon request. Some of the results were presented at the 89th meeting of the Acoustical Society of America in Austin, Texas (Repp, 1975).

more similar syllables (from different categories) contain at least one stimulus that is close to the category boundary. Therefore, the differences in categorization time for individual stimuli are confounded with the degree of acoustic discrepancy in between-category comparisons, and the effect on "different" reaction times could simply arise from the successive categorization and phonetic comparison of the two syllables. This explanation would also predict that the effect does not decrease with increasing SOA (Repp, 1975). This methodological objection does not apply to the "same" reaction times, since the individual stimuli contained in within-category comparisons can be properly counterbalanced (as in the experiment of Pisoni and Tash, 1974). One reason my study did not replicate theirs might have been the presentation of the two syllables to different ears, while Pisoni and Tash had presented them binaurally.

The present study asked the following questions:

1. Is the Pisoni-Tash effect obtained with syllables from a "place continuum," that is, with syllables whose acoustic differences lie in the initial formant transitions (and which are also perceived in a highly categorical fashion--see Pisoni, 1971)?

2. If so, does this effect decrease as the temporal separation between the syllables is increased?

3. Is there a difference in the magnitude of the effect when the syllables are presented to different ears rather than binaurally?

4. Does the Pisoni-Tash effect really reflect auditory comparisons between the two syllables, or does it perhaps consist in an influence of the first syllable in a pair on the categorization time of the second syllable? Entus and Bindra (1970) and Eichelman (1970), among others, have provided evidence that "same-different" reaction times and sequential effects in simple choice-reaction time are related and may reflect the same underlying processes. This was investigated here by including a condition in which the subjects had to classify the second syllable in each pair, ignoring the first syllable.

The study used a design that avoids the methodological problem with "different" reaction times discussed above. This design requires three categories on a single acoustic continuum, which is the case with a place continuum (/b/-/d/-/g/). Only four stimuli were used: /b/, $/d/_1$, $/d/_2$, and /g/. (The vocalic context, /ae/, was constant.) $/d/_1$ was acoustically closer to /b/ and $/d/_2$ was closer to /g/. The predictions were that "same" reaction times should be faster for $/d/_1$-$/d/_1$ and $/d/_2$-$/d/_2$ than for $/d/_1$-$/d/_2$ and $/d/_2$-$/d/_1$, and "different" reaction times should be faster for /b/-$/d/_2$ and /g/-$/d/_1$ (and their reverse orders) than for /b/-$/d/_1$ and /g/-$/d/_2$ (and their reverse orders). It can be seen that this design is completely balanced and therefore leads to unconfounded results for both "same" and "different" reaction times.

METHOD

Subjects

Eight paid volunteers (five women and three men) from the Haskins-Yale summer subject pool participated. Two of the men were left-handed. All had normal hearing and were relatively inexperienced.

Stimuli

Four synthetic syllables were produced on the Haskins Laboratories parallel resonance synthesizer. Two stimuli were supposedly good instances of /bae/ and /gae/, respectively, while the other two both sounded like /dae/ (cf. Repp, 1976b; the constant vowel will be omitted in referring to the stimuli). The two /d/s, $/d/_1$ and $/d/_2$, differed only in the onset frequencies of the second formant (F_2), which were 1620 and 1772 Hz, respectively. Since the steady-state vowel had its F_2 at 1620 Hz, $/d/_1$ had a flat F_2, while $/d/_2$ had a falling transition. The third formant (F_3) fell from 3026 to 2862 Hz in both /d/s. Likewise, /b/ and /g/ differed only in their F_2-transitions (starting at 1232 and 2156 Hz, respectively) and had identical F_3-transitions (starting at 2180 Hz). All syllables were of 280-msec duration, with 15 msec of prevoicing, no bursts, and a constant fundamental frequency (114 Hz).

Of the sixteen possible ordered pairs of the four syllables, /b/-/g/ and /g/-/b/ were omitted and /b/-/b/ and /g/-/g/ were duplicated instead. This resulted in an equal number of "same" and "different" pairs. Four stimulus lists were recorded. Each contained 80 syllable pairs, viz. 5 blocked replications of the 16 pairs, randomized within blocks. In the first and fourth lists, the SOA was 500 msec; in the second and third lists, the SOA was 2 sec. Each stimulus pair was preceded by a 100-msec warning buzz that came on 500 msec before the first syllable. The two syllables in a pair were recorded on separate channels. The interpair interval was 3 sec.

Procedure

Each subject participated in two 90-minute sessions on different days. The sequence of the two tasks was counterbalanced across subjects. In one session, the subject was instructed to judge whether the two syllables in a pair were the same or different by pressing the response key with the appropriate label (same-different task). In the other session, the instructions were to ignore the first syllable and to classify the second syllable as either "D" or "non-D," that is, "B or G" (classification task). The subjects were told that there were three syllables, /bae/, /dae/, and /gae/. In the classification task, they were informed that /b/ and /g/ never occurred together in a pair but that, apart from this, the first syllable provided no clue about the second syllable. The subjects were encouraged to be as fast and as accurate as possible. A practice series of 32 pairs at SOA = 500 was presented at the beginning of each session.

In each session, a subject listened to the experimental tape twice, once binaurally and once dichotically (i.e., with the warning tone and the first syllable in one ear and the second syllable in the other ear; "dichotic" is used here in the wider sense of "different--but not necessarily simultaneous--inputs to the two ears"). The sequence of the two presentation modes was counterbalanced across subjects, but it was the same in both sessions for a given subject. Which ear received the first syllable in the dichotic condition was also counterbalanced across subjects, but fixed for each individual subject.

The tape was played back from an Ampex AG-500 tape recorder through an amplifier/attenuator to Telephonics TDH-39 earphones. Playback intensity was approximately 88 dB SPL (peak deflections on a voltmeter). Dichotic and binaural presentation modes were established by means of electronic switches. Reaction times were recorded on a Hewlett-Packard 522B electronic counter, which

was started by the onset of the warning tone and stopped by depressing either response key. Appropriate constants were subsequently subtracted from all latencies, so that they were measured with reference to the onset of the second syllable in a pair. The subject used both hands for responding, one for each key. Hand-response assignment was again counterbalanced across subjects.

At the end of the second session, each subject was questioned whether he or she had noticed anything about the stimuli that had not been mentioned in the instructions, and subsequently, given that there were two different versions of one syllable, which of the three syllables this might have been. Of the eight subjects, three showed no awareness whatsoever (they also had the lowest error rates), three stated that /d/ and /g/ were more difficult to discriminate than /d/ and /b/, and the remaining two claimed hearing /blae/ on occasion (these two had the highest error rates). No subject guessed correctly that two /d/s were involved; all guesses were either "two /b/s" or "two /g/s."

The first step in the data analysis was to calculate the medians of the reaction times for the five replications of each stimulus pair in each list, omitting errors. Further analysis was in terms of the means of these medians.

RESULTS

Errors

Since latencies cannot be fully understood without taking the error pattern into consideration, the errors shall be presented first. There was great individual variation: average error rates ranged from 2.0 to 17.7 percent. As pointed out above, they were positively related to the degree of awareness the subject had of the presence of four stimuli. However, no subject made consistent misclassifications or misjudgments of certain stimuli; several changed their error trends during the course of a session.

The overall error rates in the two tasks were similar (same-different: 9.3 percent; classification: 9.2 percent). There was a tendency to commit more errors at the shorter SOA (10.2 percent) than at the longer one (8.3 percent). The most striking difference was between the dichotic and binaural conditions, with almost twice as many errors in the former (11.8 percent) than in the latter (6.7 percent). As might be expected, this difference was more pronounced in the same-different task, but it was also present in the classification task.

In the classification task, /d/ stimuli were misclassified more often than /b/ and /g/ stimuli (14.2 vs. 4.2 percent). Most of the errors on /b/ and /g/ were probably due to inattention and/or hand-response confusions that were not separately identified in this study (i.e., subjects were not asked to "correct" their own errors). $/d/_1$ was misclassified more often than $/d/_2$ (18.8 vs. 9.6 percent). Misclassifications of /d/ as /b/ or as /g/ were not distinguishable in this task, but it seems likely that $/d/_1$ was mostly confused with /b/, and $/d/_2$ with /g/. The nature of the preceding stimulus seemed not to make any difference.

In the same-different task, two interactions similar to those predicted for the latencies were expected, since errors and latencies tend to be positively correlated in same-different tasks. Incorrect "same" judgments should have been more frequent in $/d/_1$-/b/ and $/d/_2$-/g/ (and reverse) pairs than in $/d/_1$-/g/ and

/d/$_2$-/b/ (and reverse) pairs, and incorrect "different" judgments should have been more frequent in /d/$_1$-/d/$_2$ (and reverse) pairs than in /d/$_1$-/d/$_1$ and /d/$_2$-/d/$_2$ pairs. Both trends were present but not very pronounced (13.3 vs. 9.9 percent, and 9.5 vs. 8.1 percent, respectively). Most surprising was the fact that /b/-/g/ (and reverse) pairs did not show a substantially lower error rate than other pairs (9.2 percent). Clearly, then, the same-different judgment errors could not be predicted from the classification errors, which were more than three times higher for /d/ stimuli than for /b/ and /g/ stimuli. This indicated either that the two stimuli in a pair were matched before complete classification, or that the classification of the second syllable was not independent of the preceding syllable. No such dependence was evident in the classification errors, however.

Latencies

It was anticipated that the latencies of subjects with high and low error rates might have to be considered separately, because of the positive correlation between errors and latencies that is usually found. However, this proved to be unnecessary, since the results were completely negative, overall and for each individual subject. While some effects may not have reached significance because of the small number of subjects, the differences of principal interest were clearly not obtained.

Consider first the same-different task. The results for "same" judgments are shown in the first three columns of Table 1. In three of the four conditions, the predicted interaction (the difference between the second and third columns) was in the expected direction (positive) but small; in the fourth condition, binaural at SOA = 2000, it was in the opposite direction. No difference reached significance, and no individual subject showed a clear pattern. The more consistent trend toward longer reaction times at SOA = 2000 than at SOA = 500 also fell short of significance.

TABLE 1: Reaction times in the same-different task. (Note: the plus sign indicates that the reverse order of the stimuli is included.)

		"Same"			"Different"	
Mode	SOA	/b/+/b/ /g/+/g/	/d/$_1$+/d/$_1$ /d/$_2$+/d/$_2$	/d/$_1$+/d/$_2$	/b/+/d/$_1$ /g/+/d/$_2$	/b/+/d/$_2$ /g/+/d/$_1$
Dichotic	500	523	536	552	547	547
	2000	560	565	581	582	589
Binaural	500	526	521	542	562	564
	2000	550	569	546	586	581

The last two columns of Table 1 show the "different" latencies. Here, it was predicted that the latencies in column 4 would be shorter than those in column 5. Clearly, there was no difference at all. The only consistent tendency seems to be again longer latencies at the longer SOA, but it did not reach significance. It will also be noted that "same" latencies were somewhat faster than "different" latencies, a difference that is commonly found and was not tested for significance.

There were two effects that did reach significance: the Mode × Order and Mode × "B vs. G" interactions ($p < .01$ and $p < .05$, respectively). They are shown in Table 2.

TABLE 2: Two interactions in the same-different task. (Note: the dash indicates a specific order of the two stimuli in a pair. /d/ implies both $/d/_1$ and $/d/_2$.)

Mode	/b/-/d/	/g/-/d/	/d/-/b/	/d/-/g/
Dichotic	564	553	584	566
Binaural	567	596	554	574

It can be seen that, in the dichotic condition, pairs in which /d/ occurred first tended to have longer "different" latencies than pairs in which /d/ occurred second, and pairs containing /b/ tended to have longer latencies than pairs containing /g/. The opposite was true in the binaural condition. These effects are difficult to interpret.

We turn now to the classification condition. The results for those syllables that were preceded by a syllable from the same category are shown in the first three columns of Table 3.

TABLE 3: Reaction times in the classification task.

Mode	SOA	/b/-/b/ /g/-/g/	$/d/_1$-$/d/_1$ $/d/_2$-$/d/_2$	$/d/_1$-$/d/_2$ $/d/_2$-$/d/_1$
Dichotic	500	543	547	518
	2000	600	570	594
Binaural	500	533	544	534
	2000	562	532	537

Mode	SOA	/b/-$/d/_1$ /g/-$/d/_2$	/b/-$/d/_2$ /g/-$/d/_1$	$/d/_1$-/b/ $/d/_2$-/g/	$/d/_2$-/b/ $/d/_1$-/g/
Dichotic	500	567	552	505	504
	2000	619	608	595	570
Binaural	500	564	563	527	535
	2000	578	587	530	540

Again, there is no clear evidence for the expected effect (faster latencies in column 2 than in column 3). In the dichotic mode, there was a notable tendency to be slower at SOA = 2000 (not significant), which indicates that the preceding syllable was not completely ignored. The results for syllables preceded by a syllable from a different category are shown in the remaining columns of Table 3. Again, there is no obvious difference between columns 4 and 5, and

columns 6 and 7. However, /b/ and /g/ classification was faster than /d/ classification, and the latencies were again longer at SOA = 2000. No effect reached significance. A facilitating effect of a preceding stimulus from the same category may be noted, but only for /d/ classification.

DISCUSSION

This experiment provided no evidence for the availability of auditory information in the comparison of syllables from a "place continuum." Although only eight subjects were tested, their results make it quite unlikely that any significant effects would emerge in a larger sample, except for the trivial findings that latencies increase with SOA and that "same" latencies are faster than "different" latencies. Note that, although the data for "same" latencies in Table 1 may be suggestive of a small effect, no individual subject showed a clear pattern of results, despite reasonably stable data (10 replications of each stimulus pair).

Of course, it is entirely possible that the results of Pisoni and Tash (1974) pertain only to differences in VOT, a temporal variable, while differences in formant transitions are not retained in auditory memory. However, a study conducted independently at about the same time as the present experiment by Eimas and Miller (1975) did find a positive effect.

Their study is the more remarkable because it used stimuli from the identical place continuum, originally prepared at Haskins Laboratories by David Pisoni (see Pisoni, 1971). (The present /b/, /$d/_1$, and /$d/_2$ were their stimuli 1, 6, and 8, respectively--see their Table 1. Their continuum did not include /g/ stimuli.) They used a design similar to that of Pisoni and Tash (1974), counterbalanced for "same" pairs but not for "different" pairs. Miller and Eimas were aware of the alternative explanation for effects on "different" reaction times and emphasized the comparison of "same" reaction times for identical and nonidentical pairs. There were three SOAs (310, 460, and 1000 msec) that were randomized. At the intermediate SOA, which approximates the shorter SOA in the present study, they found a 44-msec difference in "same" reaction times and a 73-msec difference in "different" reaction times, both in the predicted direction. Moreover, the effect on "same" reaction times, but not that on "different" reaction times, decreased as SOA increased. This provides convincing evidence for the involvement of some auditory memory at short SOAs and for its decay over time. It also suggests that the effect on "different" reaction times probably does not reflect auditory memory but differences in categorization time for the component stimuli.

Eimas and Miller's study is elegant and well-designed, and their results must be taken seriously. It will require further research to clarify why the present study did not obtain the same effects, in the absence of any obvious flaws in design. Of course, if the effect of "different" reaction times is due to differences in categorization time alone, no effect should have been obtained in the present balanced design because such differences cancel out. Seen in this way, this portion of the present results even supports Eimas and Miller. However, the reason for the present failure to obtain an effect of acoustic differences on "same" reaction times remains obscure.

REFERENCES

Besner, D. and M. Coltheart. (1975) Same-different judgments with words and nonwords: The differential effect of relative size. Mem. Cog. 3, 673-677.

Cole, R. A., M. Coltheart, and F. Allard. (1974) Memory of a speaker's voice: Reaction time to same- or different-voiced letters. Quart. J. Exp. Psychol. 26, 1-7.

Eichelman, W. H. (1970) Stimulus and response repetition effects for naming letters at two response-stimulus intervals. Percept. Psychophys. 7, 94-96.

Eimas, P. D. and J. L. Miller. (1975) Auditory memory and the processing of speech. In Developmental Studies of Speech Perception (Walter S. Hunter Laboratory of Psychology, Brown University, Providence, R. I.), Progress Report No. 3, pp. 117-135.

Entus, A. and D. Bindra. (1970) Common features of the "repetition" and "same-different" effects in reaction time experiments. Percept. Psychophys. 7, 143-148.

Fujisaki, H. and T. Kawashima. (1970) Some experiments on speech perception and a model for the perceptual mechanism. In Annual Report of the Engineering Research Institute (Faculty of Engineering, University of Tokyo) 29, 207-214.

Pisoni, D. B. (1971) On the nature of categorical perception of speech sounds. Unpublished Ph.D. thesis, University of Michigan.

Pisoni, D. B. (1973) Auditory and phonetic memory codes in the discrimination of consonants and vowels. Percept. Psychophys. 13, 253-260.

Pisoni, D. B. and J. Tash. (1974) Reaction times to comparisons within and across phonetic categories. Percept. Psychophys. 15, 285-290.

Posner, M. I. (1969) Abstraction and the process of recognition. In The Psychology of Learning and Motivation. Advances in Research and Theory: Vol. III, ed. by G. H. Bower and J. T. Spence. (New York: Academic Press).

Posner, M. I., S. J. Boies, W. H. Eichelman, and R. L. Taylor. (1969) Retention of visual and name codes of single letters. J. Exp. Psychol. (Monograph Suppl.) 79, 1-16.

Posner, M. I. and S. W. Keele. (1967) Decay of visual information from a single letter. Science 158, 137-139.

Posner, M. I. and R. F. Mitchell. (1967) Chronometric analysis of classification. Psychol. Rev. 74, 392-409.

Repp, B. H. (1975) Categorical perception, auditory memory, and dichotic interference: A "same"-"different" reaction time study. J. Acoust. Soc. Am., Suppl. 57, S51(A).

Repp, B. H. (1976a) Effects of fundamental frequency contrast on identification and discrimination of dichotic CV syllables at various temporal delays. Mem. Cog. 4, 75-90.

Repp, B. H. (1976b) Identification of dichotic fusions. Haskins Laboratories Status Report on Speech Research SR-45/46 (this issue).

Springer, S. P. (1973) Memory for linguistic and nonlinguistic dimensions of the same acoustic stimulus. J. Exp. Psychol. 101, 159-163.

Studdert-Kennedy, M., A. M. Liberman, and K. N. Stevens. (1963) Reaction time to synthetic stop consonants and vowels at phoneme centers and at phoneme boundaries. J. Acoust. Soc. Am. 35, 1900.

Weak Syllables in a Primitive Reading-Machine Algorithm

George Sholes

ABSTRACT

Weak syllables are syllable types in the pronouncing dictionary of the reading machine. Weakened syllables, in the output string of the machine, come either from weak dictionary syllables or from full dictionary syllables that have been subjected to gradation. In either case, weakened syllables are further subject to certain mergers and may exhibit special segmental allophones. Weakened syllables of all kinds may also condition shortening of the full syllables they immediately follow. This compression seems to come from a kind of inclusion of the weak syllable by the full syllable. It does not occur across phonological word boundaries and by this fact helps to identify phonological word boundaries in the output.

Weak syllables, in this version of mechanical American English, are a special syllable-type which, among other things, typically comes to carry the lowest level of stress and so ends up at the bottom of the prominence heap. But weak and weakened syllables are also terms involved in a number of key operations, among which are gradation, certain neutralizations, and the selection of special segmental allophones. Finally, weak syllables condition a noticeable compression of full syllables they immediately follow. The absence of such compression, when a phonological word boundary intervenes, is a strong cue for the presence of the word boundary.[1]

In Section I of this paper syllable-types in the machine will be outlined and the operations of gradation, neutralization, and allophone selection will be identified. In Section II the shortening effect of weak syllables on full syllables will be explored. Between the two sections a brief interlude will characterize the machine itself including the pronouncing dictionary and phonological string, of which weak and weakened syllable-types are parts.

---

[1]The phonological string of the machine is a hierarchical structure of segmentals, syllables, phonological words, and phonological phrases (cf. Pike, 1945, 1967). What are called phonological words here are called total contours in Pike (1945) and stress groups in Pike (1967). What are called phonological phrases here are called rhythm units in Pike (1945) and pause groups in Pike (1967). What are called weakened syllables here are among those tentatively called ballistic syllable-types in Pike (1967:368-369).

[HASKINS LABORATORIES: Status Report on Speech Research SR-45/46 (1976)]

SECTION I

Syllable Types

In the pronouncing dictionary of the machine, phonetic entries are made up of combinations of three types of syllables. Weak, as a syllable-type in the dictionary, is illustrated by the last syllable of the following print-words: "soda, city, window, Hindu, beater, beetle, bottom, cotton, rotting." The other two types of syllable in the dictionary are stressable and plain. Stressable syllables are illustrated by the first syllable of the print-words in the list just given. Plain syllables are those which never take stress (much less pitch-accent), on the one hand, and, on the other, are not subject to mergers (neutralization); nor do they condition full syllable shortening. Illustrations of plain syllables are the first syllables of "ideal" and of "psychology" and the last syllables of the verb "veto" (but not the noun) and of "telephone." In sum, plain syllables--"ideal, telephone"--will never be stressed in any text occurrence; neither will they be degraded, that is, replaced by schwa or a weak syllabic sonorant.

It is to be noted that each print-word pronunciation in the dictionary contains at least one stressable syllable and that some pronunciations contain two or more stressable syllables. Examples of multistressable print words are "sardine(s)" and"pastel(s)" (both syllables) and "intonation" and "California" (first and third syllables). In citation pronunciation, because it means end-of-phrase, the last stressable syllable in a multistressable word would normally be stressed (and get the pitch accent): "(can of) sardines," "(box of) pastels," "intonation," "California." Within a phrase, an earlier stressable syllable may be stressed: "sardine sandwich," "pastel picture," "intonation contour," "California sunshine." The number of weak or plain syllables in a dictionary pronunciation has no upper or lower limits.

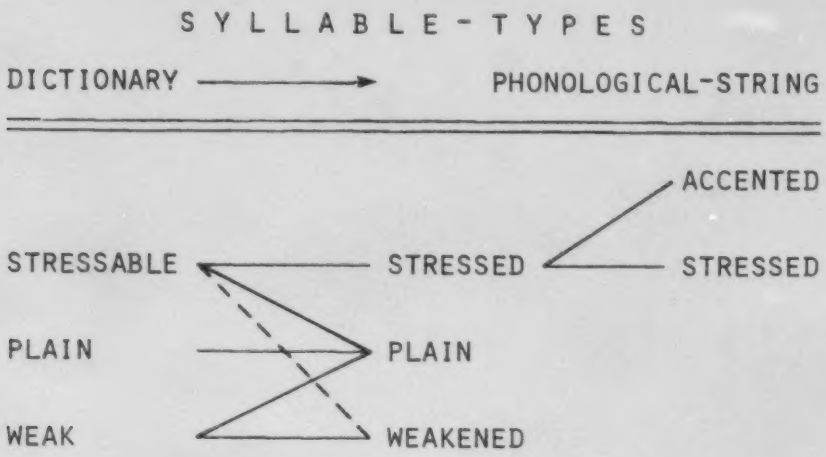
The distinction between stressable and stressed is thus one between dictionary pronunciation--stressable--and phonological string pronunciation--stressed. In the dictionary, stress is a potential of certain syllables; the stressable, a potential which may or may not be realized in some occurrence in a phonological string. A similar distinction applies to weak syllables in the dictionary and actually weakened syllables in the phonological string. By contrast, plain syllables in the dictionary carry over only into plain syllables in the phonological stress string. Figure 1 shows the possibilities.[2]

Gradation

The dashed line from stressable to weakened, which breaks a certain symmetry in Figure 1, represents the working of the operation called gradation.

---

[2] The three syllable-types in the dictionary correspond to the three stress levels posited by Newman (1946), if one moves Newman's sonorous weak in pre-heavy position to reading-machine plain. Component features that would define the four types in the phonological string could correspond with the first three suprasegmental features of Vanderslice and Ladefoged (1972): plus or minus heavy, accent, intonation. Correspondences can be made with other three- and four-way systems.



Gradable syllables in the dictionary may be realized as stressed, plain, or weakened in the phonological string. By contrast, most stressable syllables may be realized only as stressed or plain. Gradation applies to a small number of monosyllabic structural words, such as "of," "at," "do." Only some four dozen dictionary words are subject to gradation, but they are all very frequent text words. When a gradable syllable does appear in weakened form, it behaves like weakened syllables which come in the usual way from dictionary weak syllables; a syllable weakened by gradation is just like any other weakened syllable.[3]

For ease of exposition, it is useful to have a cover term for nonweak or nonweakened syllables. Full syllable will be the label that includes stressable and plain, or stressed and plain syllables.

Allophone Selection

When print words are strung together, consonant segmentals may come together at print-word boundaries. These consonant clusters may be smoothed out by reduction (dropping) or by altering component features when the syllable-type sequence over the print-word boundary is full-plus-weak. For example, the print word "miss" is stored in the dictionary with the citation pronunciation ['mɪs] and the (gradable) print word "you" with ['yuʷ]. Yet the print-word sequence "miss you," particularly in a larger context, such as "I'm going to miss

[3] See, for instance, Kenyon (1950:104-114) and Gimson (1964:239-243).

you a lot," will give the phonological string fragment ['mɪʃw̥].[4] This assembled fragment is quite similar to the string representation of the single print word "issue" ['ɪʃw̥] in the same context: "I'm going to issue a lot." It will be seen that the print-word boundary in the vicinity of the (de)graded and weakened syllables of structural words may be heavily camouflaged.

A number of single consonants have special allophones in the position between full and weak syllabics ("intervocalic position"), for example, [t, d] are flapped and [g] appears as a fricative. The special allophone is selected regardless of where the print-word (lexical) boundary falls. For instance, the fragment [meʸdn̥] can represent the first two words of "made in France," with print-word boundary on the right-hand side of the [d], or it can represent the entire word "maiden" with no print-word boundary at all abutting the [d]. Similarly, the fragment [biʸkn̥] could represent all the print-word sequences, "beacon," "bee can," "beak and," embedded in some larger context. (This is not to say that the print-word sequences cannot be distinguished, but rather that they may not be.)

Neutralization

Syllables may also be weakened--carried into the phonological string as weakened syllables--by neutralization or merger of syllable-center tambers. For example, the syllable centers of the dictionary weak syllables of "windows" and "Hindus" merge into a single tamber when those weak syllabics turn up in various nonfinal contexts, such as:

All the windows are here. 'ɔlðə'wɪndw̥z̥r̥'hɪr//

All the Hindus are here. 'ɔlðə'hɪndw̥z̥r̥'hɪr//

Whereas in various final contexts, the syllabics of these print words are quite distinct (and in the example below the dictionary weak syllables have been assembled as plain syllables):

Here are all the windows. 'hɪrɹ̥'ɔlðə'wɪn,doʷz//

Here are all the Hindus. 'hɪrɹ̥'ɔlðə'hɪn,duʷz//

In natural speech, the merged syllable [w̥] would have a tamber range overlapping part of full syllable [ʊ, uʷ] and perhaps [oʷ].[5] In sum, the allophone range of certain weakened syllabics differs from the corresponding full vowel range.

Similar contexts cue the merger of dictionary vowels [ʌ] and [ɪ]. For example, the print words "him" and "them" are indistinct in:

I can see him now.

I can see 'em now. ,aʸkn̥'siʸm̥'naʷ//

[4] A small circle below a letter has been used to indicate a weakened-syllable center: [ə y̥ w̥ r̥ l̥ m̥ n̥ ŋ̥]. Alternatively (and equivalently), the same weakened-syllable centers could be written schwa or schwa plus sonorant consonant: [ə əy əw ər əl əm ən əŋ].

[5] See, for instance, Kingdon (1969:10) and Bolinger (1963:22).

and are distinct in:

Now I can see him. 'naʷ,aʸkn̥'siʸ,ɪm//

Now I can see 'em. 'naʷ,aʸkn̥'siʸ,ʌm//

In the end, the list of weakened syllabics (vowels) in nonfinal position in the assembled phonological string is ə̥ ɤ̥ ɯ̥ ɪ̥ ɨ̥ ʊ̥ ɵ̥ ɒ̥. For this and other reasons it has from time to time been proposed that weak-syllable centers are best taken as forming a separate system apart from the larger, main system of full-syllable vowels (e.g., Hultzén, 1961; Bolinger, 1963), or that they are positional variants of the sonorant consonants (e.g., Householder, 1957). In the reading machine, however, it proves useful to have just one set of syllabics (vowels) and to have the syllable as a whole marked for its type.[6]

---

[6] The notation convention for marking syllable types is that full syllables are marked where they begin, while phonological words and phrases are marked where they end. Weak and weakened syllables are not considered to have boundaries of their own at all. By this means all distinctions of the kind "gray day" versus "grade A" and "a nice ..." versus "an ice ..." are automatically assembled. (See Jones, 1931, 1956; Lehiste, 1960; Hoard, 1966; Lee, 1970.)

However, this style of marking also requires that the syllable centers of "hot" and "heart" be written with different symbols. This is because the full vowel of "hot" may, in the assembled string, be followed by [r] and then a weakened syllable. It must still remain distinct from the full vowel of "heart" plus [r] plus weakened syllable. A test pair would be:

bas__relief vs. bar__a leaf

which can be held separate when pronounced with phonological word boundary at the points shown. When the boundary is omitted (with concomitant full-syllable compression to the left; see Section II below), the phrases are still distinct:

bas-relief ≠ bar a leaf

'bɑrə̥'liʸf// ≠ 'barə̥'liʸf//

Similarly, with phonological word boundary omitted:

Ma renewed ≠ mar a nude

'mɑrə̥'nʷd// ≠ 'marə̥'nuʷd//

and also:

paw repair ≠ pour a pair

'pɔrə̥,pɛr// ≠ 'porə̥,pɛr//

It is nonetheless possible to write the syllable center of "bird" either as a unit--[ɚ]--or as a sequence of wedge plus [r]--[ʌr]--with no contrastive difference. Full-syllable wedge will never otherwise be followed by [r] in

MACHINE INTERLUDE

With this much of a sketch of weak syllables and weak syllable operations, the reading machine itself can be characterized in general terms. It is an algorithm and a machine in the sense that it is a series of computer programs. It reads in the sense that it, together with the hardware attached to it, converts strings of print representations into an acoustic signal that is a simulation of speech. Finally, it is primitive in that a human editor is asked to intervene at one point to add information that is not available automatically.[7]

Schematically, the machine moves from print text to synthetic speech in two large steps, as shown in Figure 2. First, the print text is turned into a phonological string; then the phonological string is converted into parameter frames that drive an electronic synthesizer, the output of which is an audio signal that can be heard as speech.

The first step converts the print text into a phonological string. This involves chunking the print text up into print words, then replacing the print words by their dictionary pronunciations, and then reassembling the text. At the end of this first step, the text appears in a phonetic notation where originally it stood in ordinary English spelling.

Reassembling the text after the dictionary look-up is a procedure of some complication. The vowel mergers and consonantal simplifications suggested in Section I above are an important part of reassembly. The dictionary look-up, by contrast, is quite simple. The dictionary is presented with an orthography, such as "cat," whereupon it returns ['kæt] plus the tag for open-class words. In this way the dictionary provides the segmental phonemes and the basic syllable structure of the phonological string. The rest is up to the editor. He marks for phonological words and phrases; and, since these carry the intonation, the intonation. The editor is thus standing in for what appears to be a syntactic, semantic analysis of the print text. He is also carrying out certain independent phonological decisions.

---

this kind of American English. Schwa plus [r] may occur in weakened syllable at print-word boundary joints. When this happens, schwa plus [r] will not contrast with syllabic [r] in a weakened syllable. A test pair, with phonological word boundary included, would be:

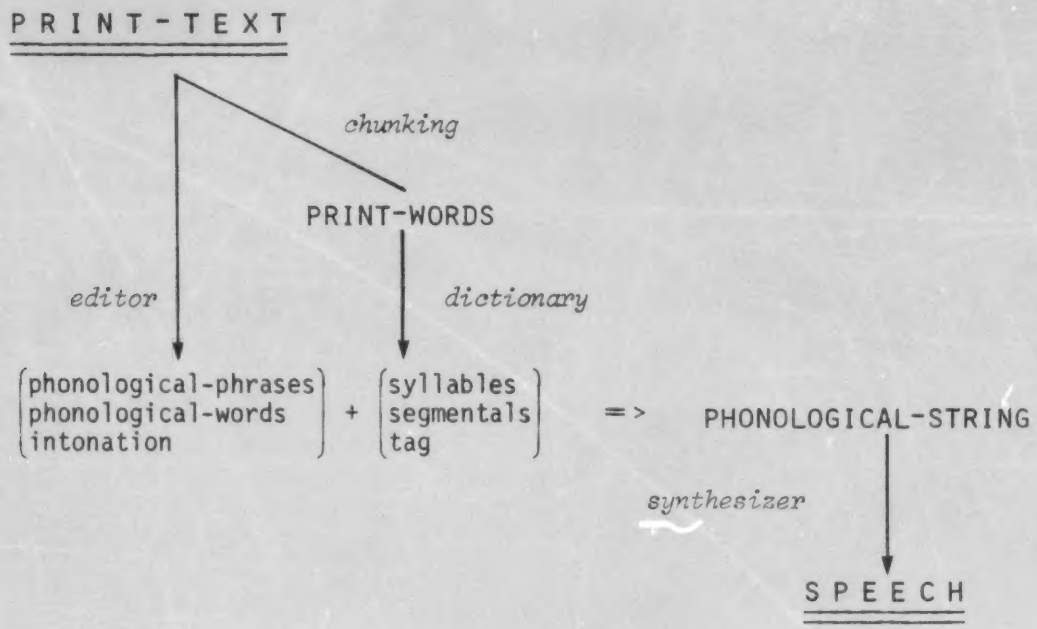
rows__are applied vs. Rosa__replied

When the boundary is omitted, the two phrases fall together and are indistinct:

'roʷzrə'plaʸd// = 'roʷzərə'plaʸd//

and in other such instances, sequences of weakened schwa plus sonorant are taken as equivalent to the syllabic sonorant alone.

[7] This characterization of the machine is not only general, it is idealized. In particular, the introduction of the editor can be taken as an expository device.



SECTION II

This section outlines an operation called compression, full-syllable compression, and it is an adjustment of durations. The units to be adjusted are full syllables, both stressed and unstressed, and the essential context for the adjustment is provided by weak syllables and phonological-word boundaries.[8]

Other things being equal, the most powerful of the interdepending cues for prominence is generally taken to be literal length: duration in time (Fry, 1970). Compression has the curious effect of making a full syllable salient by shortening its duration. The most complete description of this effect has been given by Bolinger (1963, 1965).

Consider a phrase consisting entirely of full syllables, that is, devoid of weakened syllables:

'YOU ,MAKE 'BILL ,LOOK 'GOOD //

It is generally possible to insert a weakened syllable into such a phrase with absolutely no increase in overall phrase duration. In fact, the new phrase is just as long as the original. The definite article "the" will do for insertion. It gives:

[8] What are called phonological-word boundaries here are called intonation breaks in Pike (1945). See also the discussion of Solutions A, B, and C in Pike (1967:405-409).

'YOU ,MAKE THE 'BILL ,LOOK 'GOOD //

The indefinite article and certain possessives, all as weakened syllables, do the same:

'YOU ,MAKE A / HER 'BILL ,LOOK 'GOOD //

Inserting a full syllable rather than a weak syllable does not give the same result. The phrase becomes not only longer in segmentals and syllables, it also becomes longer in total duration. The demonstrative "that" will do for full-syllable insertion. It gives:

,YOU 'MAKE ,THAT 'BILL ,LOOK 'GOOD //

When a weak syllable is inserted, something in the original phrase is compressed to make room for it. When a full syllable is inserted, this compression does not occur. What gets compressed when a weak syllable is inserted is the full syllable to the left of the weak syllable. In these examples, this is the print word "make": it is compressed in the fragments: "make the bill, make 'er bill, make a bill"; "make" stands at its normal length in the fragments: "make bill, make that."

Bolinger is at pains to point out that compression or its absence is independent of I(mmediate) C(onstituent)-cuts. The articles, demonstrative and possessives go syntactically with the next item to the right, the print word "bill": "a bill, the bill, that bill, her bill." As weak (and then weakened) syllables, they nonetheless compress the syllable to the left, "make." In short, compression is determined phonologically rather than syntactically.

Compression is obligatory in the sense that failure to compress a full syllable in this context tends to give a stage (stereotyped) Scandinavian accent, and pronunciation guides intended for Scandinavian learners of English often explicitly point out this potential stumbling point (e.g., Lewis, 1969:50-51). Full-syllable compression is obviously no language universal, and this suggests that it is not even a universal for languages that have stressed syllables, as do the Scandinavian.

By way of parenthesis, it is worth noting a possible articulatory explanation for full-syllable compression. Ladefoged (1962), attempting to correlate intercostal muscle activity with Stetson's (1951) chest-pulses, noted that certain syllable sequences may be articulated on a single burst of intercostal activity, even though the usual pairing is one chest-pulse/one syllable. He cites the word "pity" as an example, and the word "doddered" in his Figure 3 appears to have been articulated this same way.

To put is metaphorically, a full syllable in English attempts to include an immediately following weak syllable, include it in the same production gesture. There is, perhaps, a parallel with syllable-closing consonants which are also not in their most natural place at the end of a syllable. Consonants naturally begin syllables. In this sense, both syllable-final consonants and included weak syllables would be unnatural phonological structures, and of course both shorten the segmental substance that precedes "in the same syllable."

What is the magnitude of compression? Lehiste (1971) has published measurements in phrase-final position, that is, where compression is combined with phrase-final length adjustments (and those of intonation as well). She compared pairs such as "stead," a full syllable, with "steady," full-plus-weak. In this position, with such pairs, the single syllable actually averages out longer in duration than the whole compressed sequence. Not all components were equally compressible. The full vowel is most amenable to compression. Differences between regular and compressed vowel lengths are somewhat greater than two to one. The leading consonants are most resilient, though nonetheless affected. Every element in the compressed syllable is compressed to some degree.

Bolinger (1963) maintained that compression is independent of IC-cuts, independent of the syntax. In a British tradition, compression is treated as a correlation between the lexicon and the phonology. Abercrombie (1965) has given an exposition from this point of view. In the R(eceived) P(ronunciation) of British English, he notes (or perhaps declares--see Uldall, 1966, 1971) that the spacings between stressed-syllable onsets are "of (approximately) even length": RP stresses are isochronous. Yet given the roughly constant durations between stressed onsets, the included segmental meterial may be divided over the available time in different ways. Here he gives the classical contrast:

take Grey__to London vs. take Greater__London

In the phrase on the left, Abercrombie stated that the relative lengths of the syllables "Grey" and "to" are on the order of two to one, whereas in the single word "greater" the relative syllable lengths are on the order of one to one. For a comparable contrast with the segmentals of American English, there is:

the rush__and turmoil vs. the Russian__turmoil

In sum, full-syllable compression on the left-hand side of these contrasting pairs has been blocked by an immediately following word boundary. So an effective cue for the presence of this word boundary would be the sequence full plus weak syllable with an uncompressed full syllable.

Abercrombie wanted to relate (what is here called) compression to the lexical composition of the phrase. Certain structural words (proclitics in the examples above: "to," "and") are not independent words at all: they merge phonologically into their neighbors. But this way of looking at things as lexically determined, apparently, leads to overlooking yet a third possible way of distributing the same segmental material between two stressed onsets, to wit: with no included phonological word boundary at all.

The contrast of presence versus absence of phonological word boundary between two stressed onsets is demonstrated by Pike (1945:37, 1967:385) with two versions of the print phrase "a book of stories":

a book__of stories vs. a book of stories

Since Pike actually recorded these examples when the earlier book appeared, it is possible to measure his segmental durations. The difference in compression is as clear to the tape measure as it is to the ear. The full vowel of "book" followed by the boundary is about twice as long as the same full vowel followed immediately by the weak syllable "of." But the upshot of this is that

the absolute durations between stressed onsets in these two versions of "a book of stories" are distinctly different. At this level of detail, at least, English is not literally isochronous. In fact, a phonological word boundary gives what Householder (1957) calls "a significant rhythm break," and if that is so, we would expect the different overall durations we do indeed find.

So a third version of the Abercrombie and American examples is possible, this time without any included phonological word boundary, and it will be not only shorter in total duration, but lexically ambiguous as well:

take Grey to London = take Greater London

'teʸk'greʸtə'lʌndn// = ˌteʸk'greʸtə'lʌndn//

and

the rush and turmoil = the Russian turmoil

ðə'rʌʃn'tɚˌmoʸl// = ðə'rʌʃn'tɚˌmoʸl//

I suspect this is the usual way of saying these phrases when the print words "greater" and "Russian" are used, despite the ambiguity.

Now to these versions can be immediately added yet a fourth in which the weak syllable previously included is left out. Over the fragment of interest, we will now have stressed-plus-stressed, where before we had stressed-plus-weakened-plus-stressed. Some of these truncations will be nonsense sequences, but no matter:

take Grey__London

the rush__turmoil

a book__stories

The uncompressed syllables "Grey," "rush," "book" followed by phonological-word boundary here are quite comparable in length to their other occurrence followed by phonological-word boundary:

take Grey__to London

the rush__and turmoil

a book__of stories

To put it another way, when compression is blocked by a phonological-word boundary, the ongoing calculations for segmental durations would be caught up to that point: there do not seem to be durational dependencies of this kind running over the phonological-word boundary.[9]

---

[9] Phonological-word boundaries are independent of lexical word boundaries, though they frequently coincide. It is to be noted that a phonological-word boundary may appear in the middle of a single lexical item, provided the item

SUMMARY

Pronunciations from a dictionary look-up on a print text are reassembled into a phonological string which is then converted into synthetic speech. The phonological string is a hierarchical structure based on segmental phonemes which are grouped into syllables, phonological words, and phonological phrases by boundary marks inserted among the segmentals. Full syllables are marked where they begin; words and phrases, where they end. Weak syllables are taken to have no inherent boundaries at all. They may be "included" in adjacent full syllables by effects of compression and neutralization which simultaneously give the including phonological-word characteristic features of its prominence silhouette.

REFERENCES

Abercrombie, D. (1965) Syllable quantity and enclitics in English. In Studies in Phonetics and Linguistics. (London: Oxford University Press), pp. 26-34.

Berger, M. D. (1955) Vowel distribution and accentual prominence in modern English. Word 11, 361-376.

Bolinger, D. L. (1963) Length, vowel, juncture. Linguistics 1, 5-29.

Bolinger, D. L. (1965) Pitch accent and sentence rhythm. In Forms of English. (Cambridge, Mass.: Harvard University Press), pp. 139-180.

Fry, D. B. (1970) Speech recognition and perception. In New Horizons in Linguistics, ed. by J. Lyons. (Harmondsworth, England: Penguin), pp. 29-52.

Gimson, A. C. (1964) An Introduction to the Pronunciation of English. (London: Edward Arnold).

---

is realized with two stressed syllables. Any multistressable word will lend itself to this kind of realization and no more so than in ultracareful citation form. Thus we have double-stressed versions, with included phonological-word boundary, of "sardine" and "absolutely":

'sar_'diʸn// 'æbsə_'luʷtly//

and double-stressed versions without phonological-word boundary:

'sar'diʸn// 'æbsə'luʷtly//

The most usual versions retain only the last dictionary stress:

ˌsar'diʸn// ˌæbsə'luʷtly//

(See Pike, 1945:77.)

Berger (1955) notes several examples, particularly from advertising and comic strips, where this incipient ambiguity among print words and print phrases has been exploited: "Chip 'n Dale, Etta Kett, K-9 Corps," etc. A phonological-word boundary is presumably more likely than not to correspond to a lexical boundary, just as a consonant is more likely to begin a syllable than is a vowel. Absolutely, however, the occurrence of a consonant does not establish a syllable boundary and the occurrence of a phonological-word boundary does not establish lexical boundary. In this sense the phonology is independent of the lexicon, though closely related to it.

Hoard, J. E. (1966) Juncture and syllable structure in English. Phonetica 15, 96-109.
Householder, F. W. (1957) Accent, juncture, intonation, and my grandfather's reader. Word 13, 234-245
Hultzén, L. S. (1961) System status of obscured vowels in English. Language 37, 565-569.
Jones, D. (1931) The word as a phonetic entity. Le Maître Phonétique 34, 60-65.
Jones, D. (1956) The hyphen as a phonetic sign. Z. Phonetik 9, 99-107.
Kenyon, J. S. (1950) American Pronunciation. (Ann Arbor: George Wahr).
Kingdon, R. (1969) Grammar of Spoken English, ed. by H. E. Palmer and F. G. Blandford. (Cambridge, England: Heffer), vol. 10.
Ladefoged, P. (1962) Sub-glottal activity during speech. In Proceedings of the Fourth International Congress of Phonetic Sciences, ed. by A. Sovijärvi and P. Aalto. (The Hague: Mouton), pp. 73-91.
Lee, W. R. (1970) Noticing word-boundaries. In Proceedings of the Sixth International Congress of Phonetic Sciences, ed. by B. Hála, M. Romportl, and P. Janota. (Prague: Academia), pp. 535-538.
Lehiste, I. (1960) An acoustic-phonetic study of internal open juncture. Phonetica, Suppl. 5.
Lehiste, I. (1971) Temporal Organization of Spoken Language, ed. by L. L. Hammerich, R. Jacobson, and E. Zwirner. (Copenhagen: Akademisk Forlag), pp. 159-169.
Lewis, J. W. (1969) Guide to English Pronunciation. (Oslo: Universitets-forlaget).
Newman, S. S. (1946) On the stress system of English. Word 2, 171-187.
Pike, K. L. (1945) The Intonation of American English. (Ann Arbor: University of Michigan Press).
Pike, K. L. (1967) Higher-layered units of the manifestation mode of the utterance (including the syllable, stress group and juncture). In Language in Relation to a Unified Theory of the Structure of Human Behavior, 2d ed. (The Hague: Mouton), chap. 9, pp. 364-432.
Stetson, R. H. (1951) Motor Phonetics, 2d ed. (Amsterdam: North Holland).
Uldall, E. T. (1966) English RP. Le Maître Phonétique 126, 34.
Uldall, E. T. (1971) Isochronous stress in R.P. In Form and Substance, ed by L. L. Hammerich, R. Jacobson, and E. Zwirner. (Copenhagen: Akademisk Forlag), pp. 205-210.
Vanderslice, R. and P. Ladefoged. (1972) Binary suprasegmental features and transformational word-accentuation rules. Language 48, 819-838.

Control of Fundamental Frequency, Intensity, and Register of Phonation*

Thomas Baer,[+] Thomas Gay,[+] and Seiji Niimi[++]

ABSTRACT

Electromyographic activity of several intrinsic and extrinsic laryngeal muscles was recorded as untrained singers produced systematic changes in fundamental frequency (F_0), intensity, and register of phonation. For one subject, subglottal pressure was recorded simultaneously. Cricothyroid muscle activity varied most consistently with F_0 over most of the range of F_0, although the activity of several other muscles was also related to F_0. Vocalis muscle activity varied most consistently with the shift between chest and falsetto registers. Subglottal pressure varied consistently with changes in vocal intensity. Activity of the extrinsic muscles was correlated with F_0 at both the high and low extremes of the chest voice range. For at least one subject, the extrinsic muscles seemed to be solely responsible for varying F_0 at its low extreme. The activity of muscles not directly associated with the larynx also changed systematically with F_0 at the high extreme.

Recent electromyographic (EMG) studies of the control of fundamental frequency, intensity, and register of phonation have dealt with the intrinsic laryngeal muscles (e.g., Hirano, Ohala, and Vennard, 1969; Hirano, Vennard, and Ohala, 1970; Gay, Hirose, Strome, and Sawashima, 1972) or with the extrinsic muscles and subglottal pressure (Shipp and McGlone, 1971). Simultaneous recording of intrinsic and extrinsic laryngeal muscles and subglottal pressure has been reported for speech intonation (e.g., Collier, 1975) but not for singing. Thus, the purpose of this study is to reexamine the nature of the control of phonation by the intrinsic and extrinsic muscles of the larynx and by subglottal pressure.

For this study, four untrained singers produced systematic changes in fundamental frequency (F_0), intensity, and register of phonation while EMG activity was recorded using hooked-wire electrodes (Basmajian and Stecko, 1962; Hirose, 1971). For subject TB, subglottal pressure was also measured, using a cannula

---

*Paper presented at the 90th meeting of the Acoustical Society of America, San Francisco, Calif., 3-7 November 1975.

[+]Also University of Connecticut Health Center, Farmington.

[++]On leave from the University of Tokyo, Japan.

Acknowledgment: This research was supported by NIDR grants 5T22 DE00202 and DE01774.

[HASKINS LABORATORIES: Status Report on Speech Research SR-45/46 (1976)]

inserted through the cricothyroid space. Each note was produced on the syllable /bi/. Each vocal maneuver was repeated 10 to 15 times, and average results were calculated using the Haskins Laboratories EMG data processing system (Kewley-Port, 1973). This system computes average activity from several repetitions of an utterance as a function of time offset from a predetermined lineup point associated with each token.

Figure 1 shows a typical result. The subject produced one-octave arpeggios starting from a fundamental frequency in the middle of his chest-voice range. The arpeggios were performed at three different intensity levels. Average activity was calculated for each of these conditions using for lineup point the onset of voicing for the first (lowest) note (shown on the left-hand side of the figure) and also using the onset of voicing for the fourth (highest) note (shown on the right-hand side of the figure).

Average activity of the cricothyroid (CT) and vocalis (VOC) muscles was found to vary systematically with fundamental frequency, but not with intensity. (Activity of the VOC muscle was sometimes more closely correlated with F_0 than is shown in Figure 1). Subglottal pressure varied systematically with intensity (or vocal effort), but its variation with frequency was smaller and less systematic. This close correlation between subglottal pressure and intensity is qualitatively in agreement with the results of other investigators (e.g., Isshiki, 1964). We plan to investigate the relationship between subglottal pressure and fundamental frequency in more detail in the future.

Figure 2 shows similar results from subject KK--a female. Two lineup points have been used, and the results have been superimposed in their overlap region. Cricothyroid and VOC activity vary systematically with fundamental frequency but not with intensity. Activity of two extrinsic muscles, the thyrohyoid (TH) and the sternohyoid (SH), is shown. The pulsatile structure of the TH plots shows that its activity is related to the segmental gestures for producing the syllables. However, the symmetric envelope of activity centered about the second lineup point shows that its level of activity is also related to F_0. The plots of SH activity show tendencies similar to those of the TH, though they appear less dramatic in this run. The TH activity shows some differences in activity for the highest intensity condition.

In several runs, EMG activity was recorded from the inferior constrictor muscle. The electrodes were directed toward the cricopharyngeal part of the muscle, and these placements were verified using activity during swallowing. The results were inconsistent across subjects. In Figure 3, the upper plots show the inferior constrictor data corresponding to the data in Figure 2. The only increases in activity are associated with the first note and the last note. This activity appears to be related to the production of the lowest frequencies, although it could also be related to maneuvers associated with the beginning and end of the phrase. These two interpretations could be differentiated by performing descending-ascending rather than ascending-descending arpeggios in the same range, but such maneuvers were not performed. The lower plots in Figure 3 show the inferior constrictor activity corresponding to the plots in Figure 1. Here, inferior constrictor activity increases with both F_0 and intensity except for the high intensity condition, for which there is an increase of activity associated with the first and last notes. For the other two conditions, there is a decrease of activity immediately before the onset of the first note, and a small increase of activity at the end of the phrase. The meaning of these results

FIGURE 1

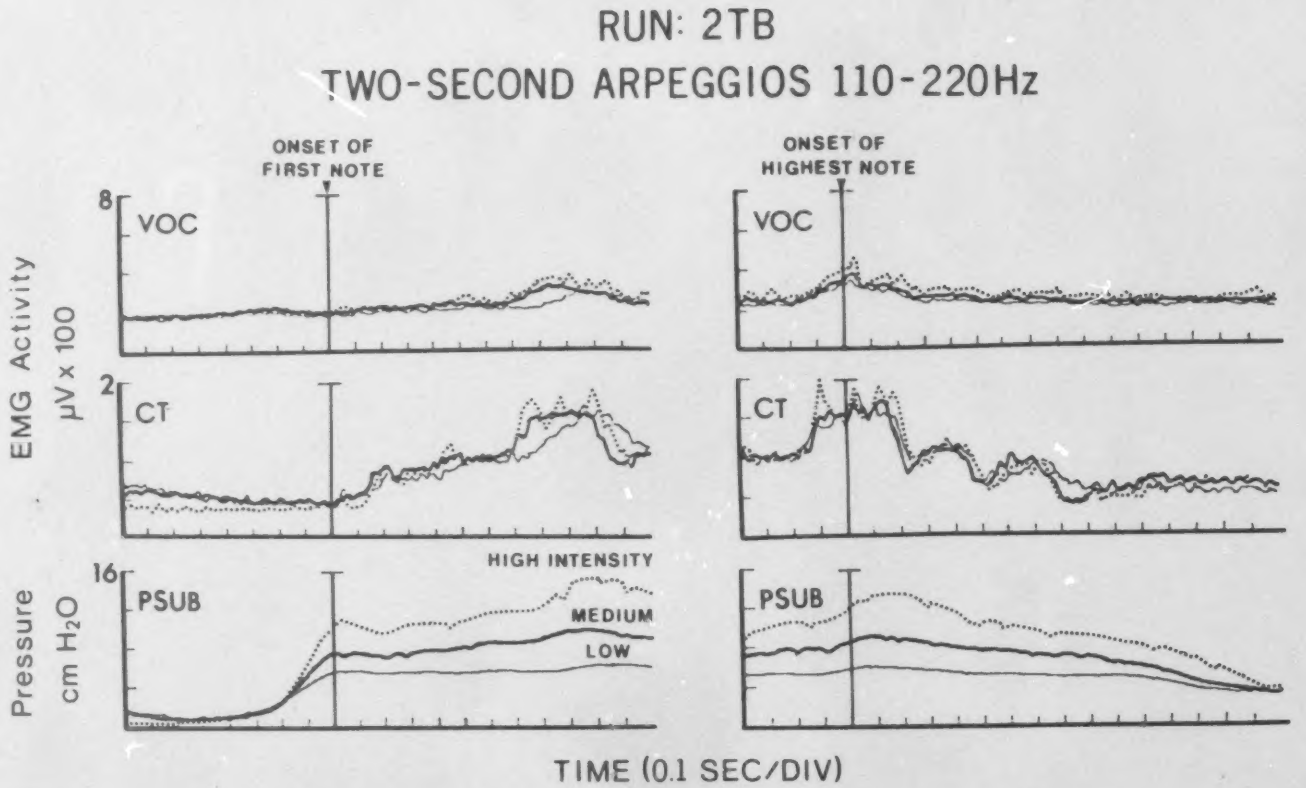


Figure 1: EMG activity and subglottal pressure during arpeggios at three intensity levels. Subject TB.

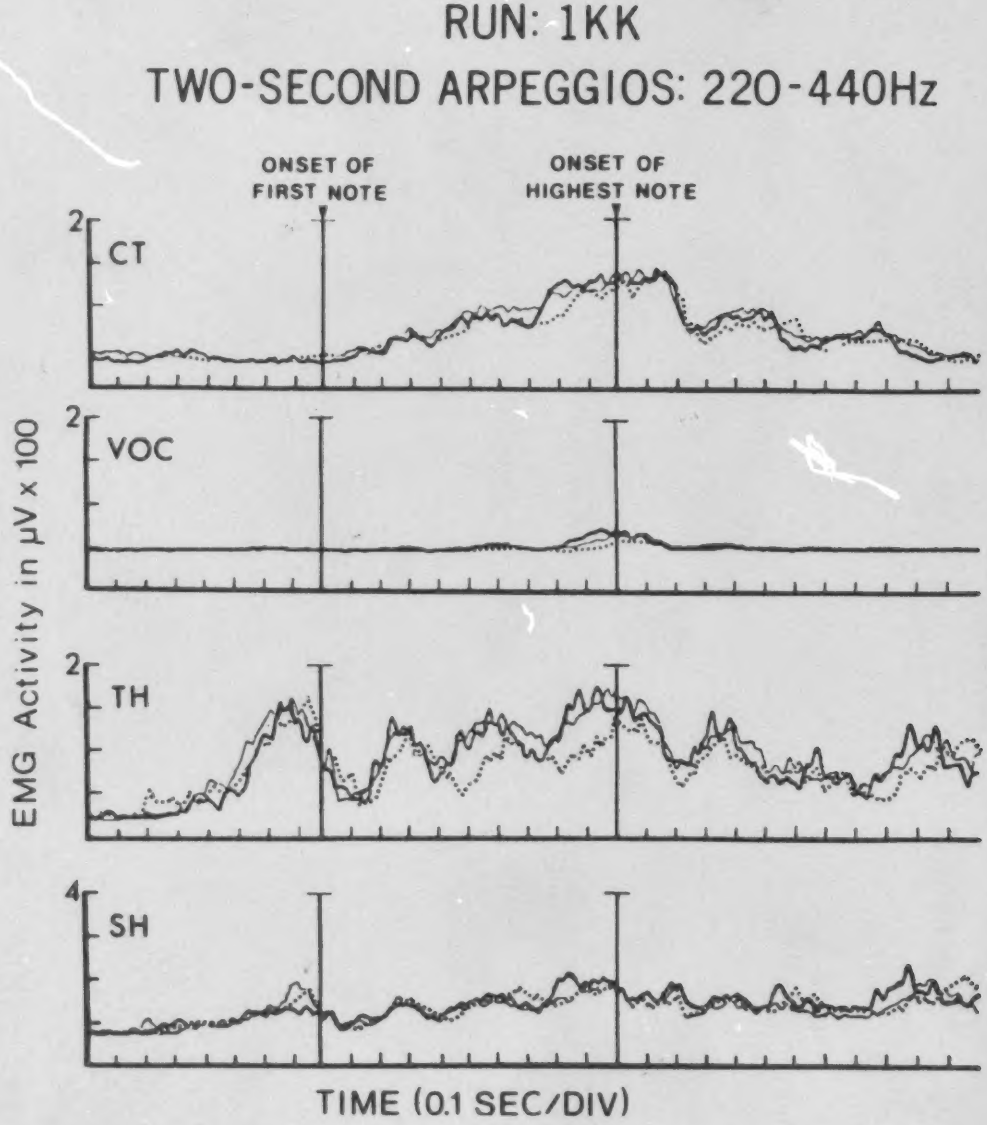


Figure 2: EMG activity of intrinsic and extrinsic muscles during arpeggios at three intensity levels. Subject KK. (Same legend as Figure 1.)

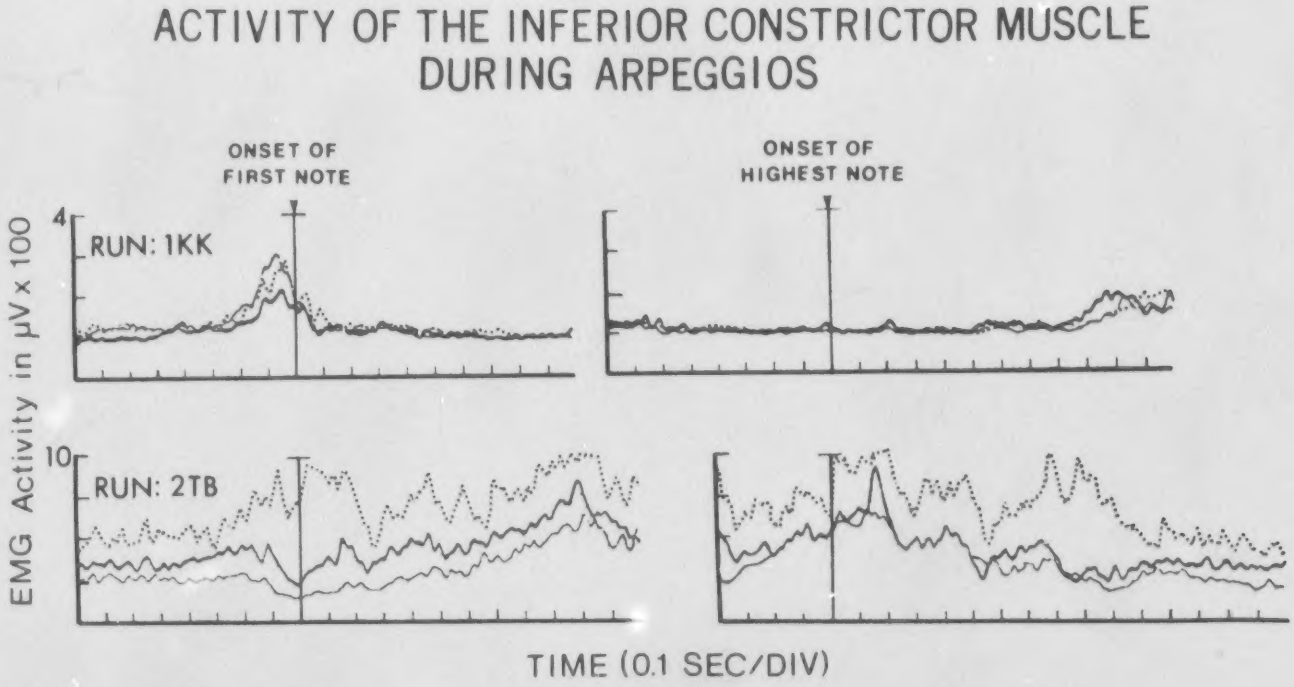


Figure 3: EMG activity of the inferior constrictor muscle during arpeggios at three intensity levels. Subjects KK and TB. (Same legend as Figure 1.)

179

is unclear, and must be further investigated with repeated insertions on the same (and other) subjects and with other vocal maneuvers.

We reconfirmed the well-known fact that extrinsic muscle activity contributes to the control of F_0 at both extremes of a subject's chest-voice range (e.g., Sonninen, 1956). Results from the low extreme are shown in Figure 4. The subject produced an ascending scale at the rate of one note per second starting at about his lowest note. Average activity of each of four laryngeal muscles and of subglottal pressure was measured for each note and plotted as a function of the fundamental frequency of the note in the figure. As the figure shows, there was no significant change in CT or VOC activity for the lowest notes, and subglottal pressure was held fairly constant throughout. However, there were clearly changes in activity of the two strap muscles--the sternothyroid (ST) and thyrohyoid (TH)--for the lowest notes. Although we had no reliable insertions into muscles other than the ones shown in Figure 4, it seems reasonable to conclude that the ST and TH, and possibly other extrinsic muscles, were responsible for producing the lowest fundamental frequencies. This result is of interest for both singing and speech, since the low extreme of the F_0 range for singing overlies the range of F_0 commonly used for speech.

At the high extreme, we examined the control of register for subject TB, who could reliably produce the same note in either chest-voice or falsetto. The results of shifting from falsetto to chest-voice on three different notes are shown in Figure 5. The subject sang the syllable /bi/, first in falsetto and then in chest-voice. The lineup point for averaging was the onset of the chest-voice note. The plots on the left-hand side of the figure show the activity of the CT and VOC muscles and of subglottal pressure. The plots on the right-hand side of the figure show activity of the inferior constrictor (IC) muscle and one strap muscle, the TH. In all cases, the activity of the VOC muscle was greater in chest-voice than in falsetto. The level of CT activity increased at the shift from falsetto to chest-voice for the 220- and 330-Hz notes, but there was only a very small increase for the 440-Hz note. The TH shows no change of activity for the lower two notes, but an increase of activity for the shift into chest-voice in the highest note. These results are consistent with the notion that the VOC muscle is most closely associated with the control of register, while the CT and strap muscles produce compensatory activity to regulate fundamental frequency. Both subglottal pressure and IC activity consistently increased during the shift from falsetto to chest-voice. The significance of this increase is difficult to assess, especially since intensity was not controlled in these maneuvers. Although the results are not shown here, equivalent results showing a general decrease of activity were obtained when the shift was made from chest-voice to falsetto.

Figure 6 shows a plot of intrinsic muscle activity at the high extreme of the chest-voice range for subject SN. The subject produced ascending scales at the rate of one note per second, and average activity for each note was plotted as a function of the F_0 of the note, as in Figure 4. In addition to the increase of CT and VOC activity with fundamental frequency, both the lateral cricoarytenoid (LCA) and posterior cricoarytenoid (PCA) muscles showed some increase of activity with fundamental frequency. Although we were not fortunate in achieving good PCA insertions, this figure shows at least one example in which there was a small but systematic increase in PCA activity at the high F_0 extreme. Such a result was reported by Gay et al. (1972), but was not evident in the data of Shipp and McGlone (1971).

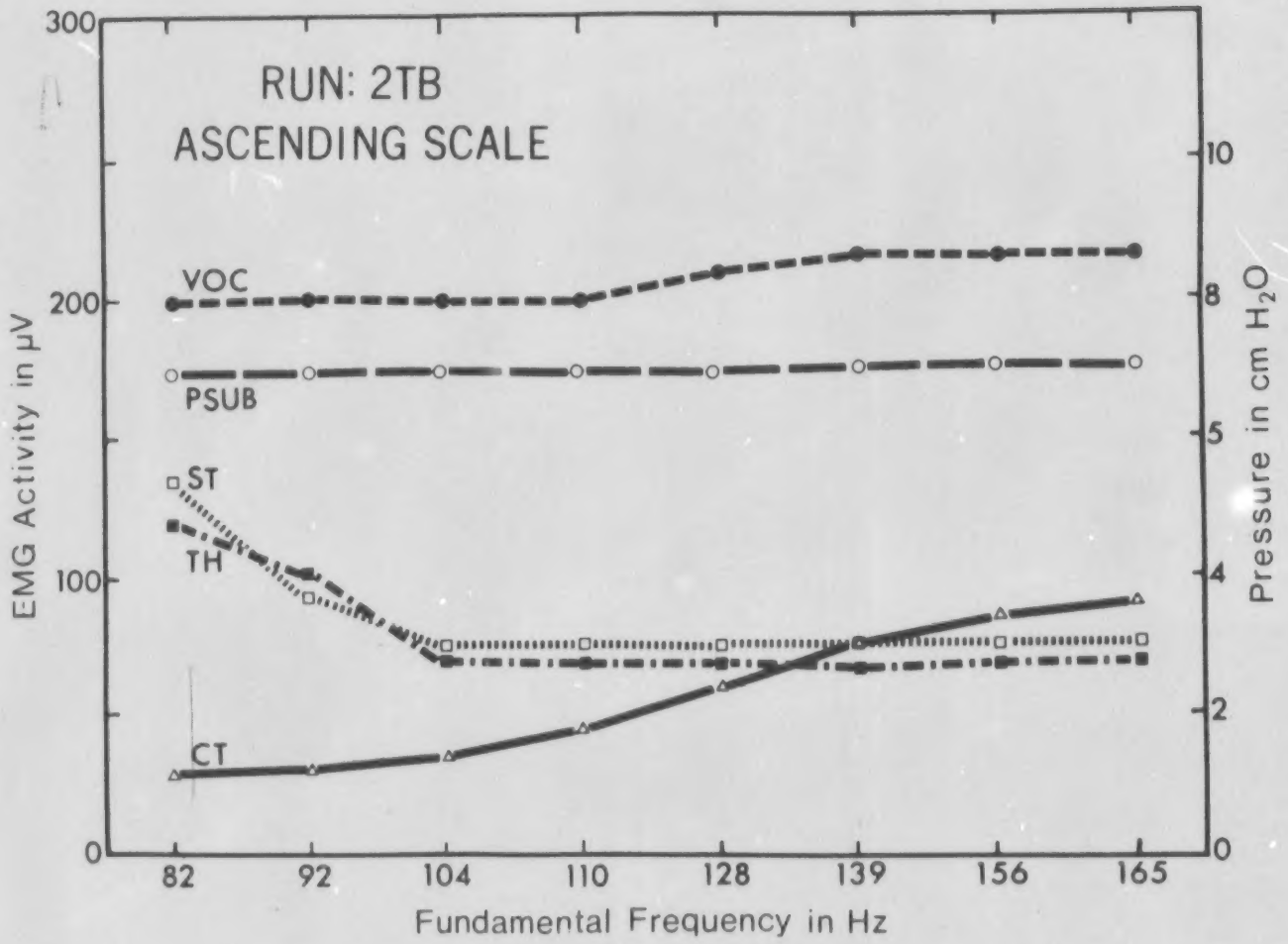


Figure 4: EMG activity and subglottal pressure versus fundamental frequency in low chest-voice. Subject TB.

181

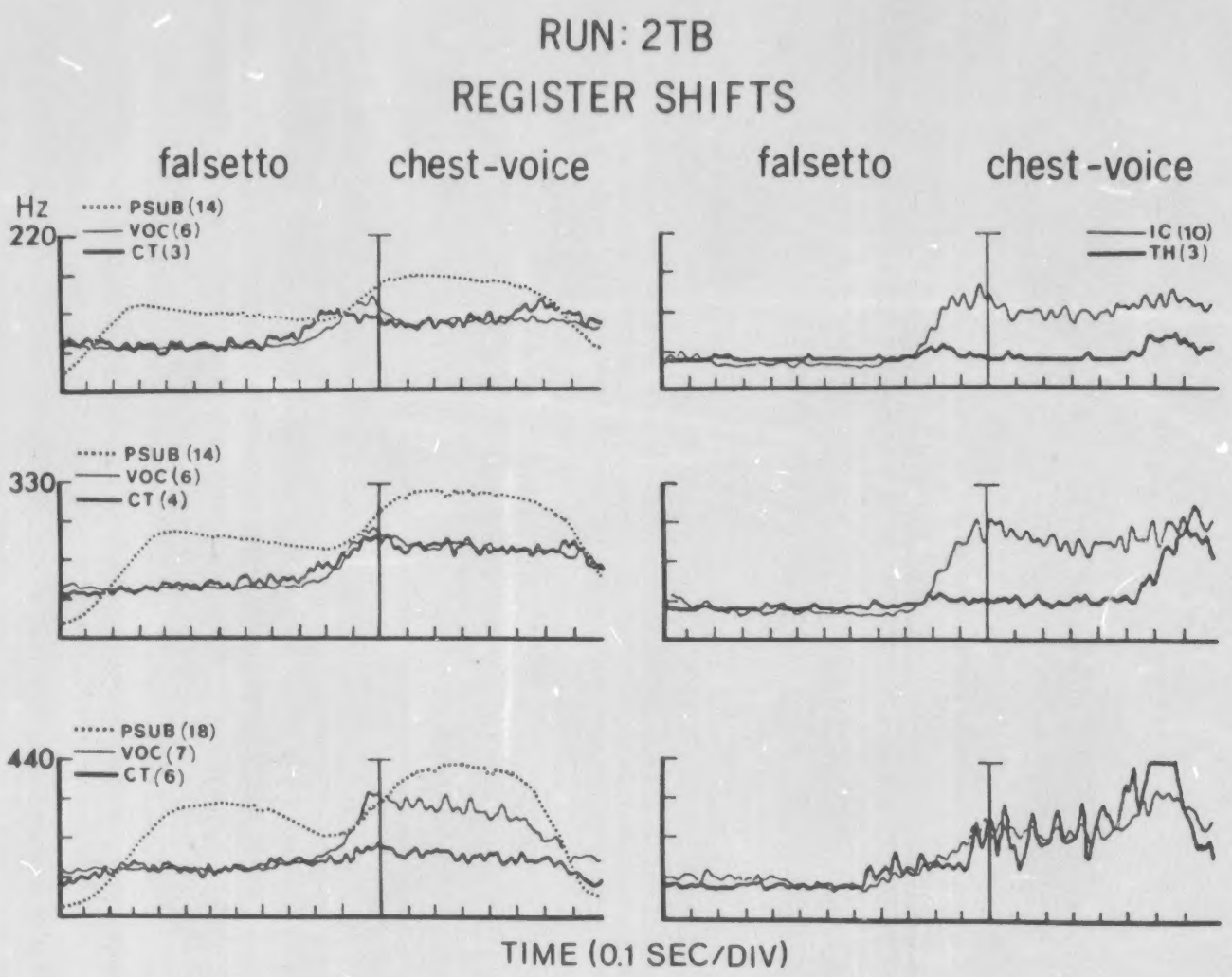


FIGURE 5

Figure 5: EMG activity and subglottal pressure during register shifts on three different notes. Subject TB. The numbers in parentheses represent the full-scale values in hundreds of microvolts (EMG) or centimeters of water (subglottal pressure).

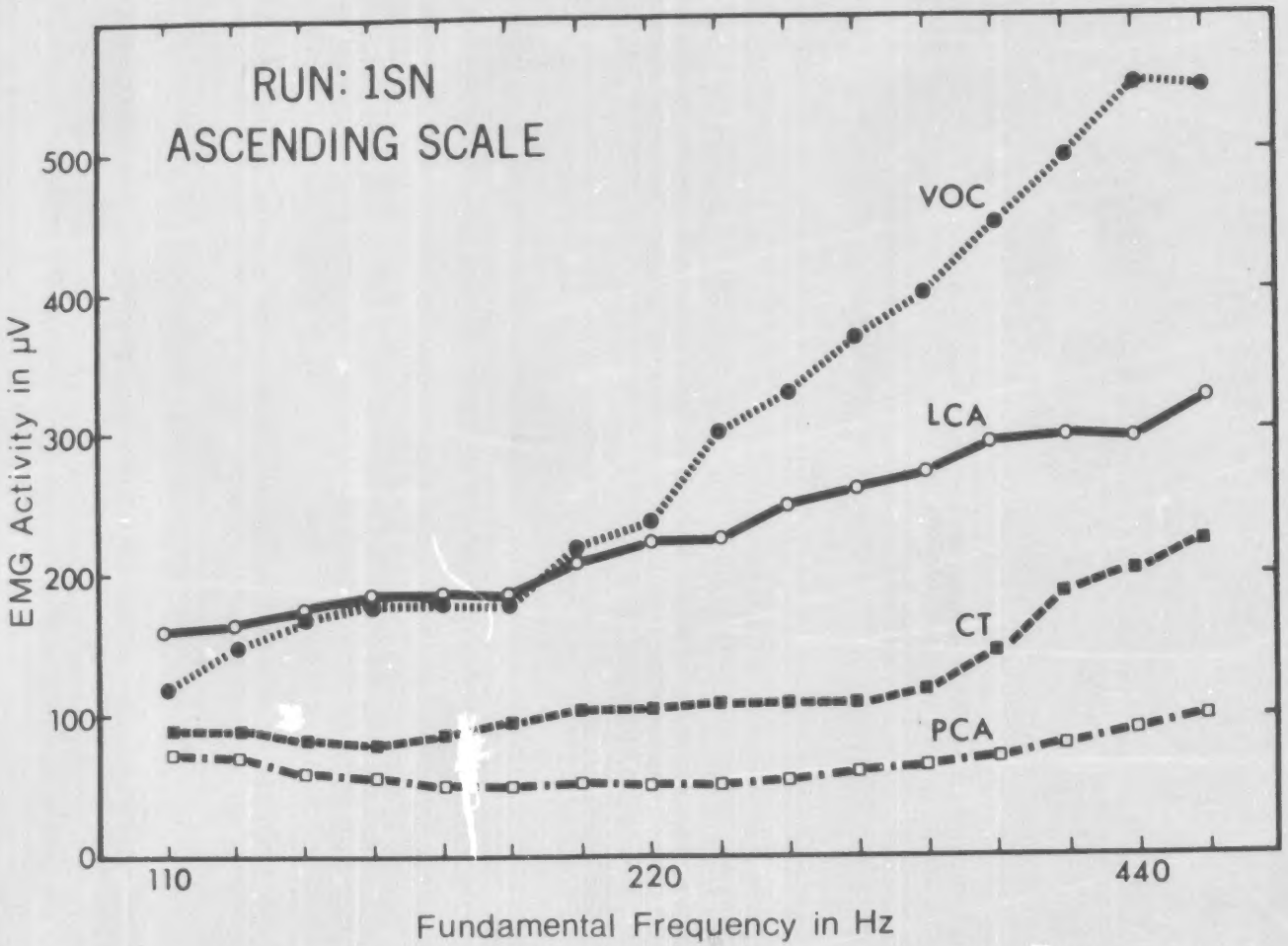


FIGURE 6

Figure 6: EMG activity of the intrinsic muscles versus fundamental frequency in high chest-voice. Subject SN.

183

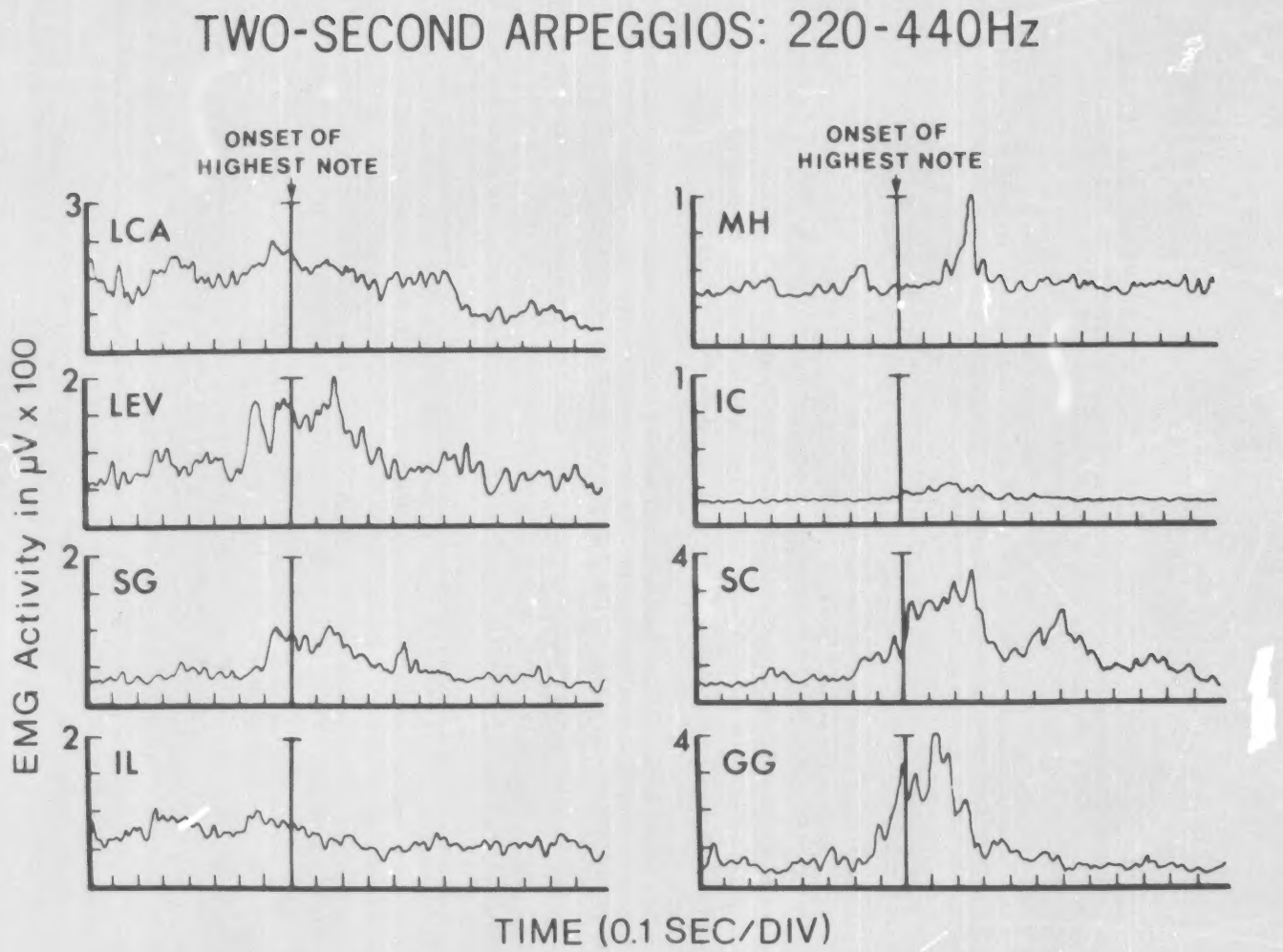


Figure 7: EMG activity of several muscles during arpeggios in high chest-voice. Subject TB.

A final point is made in Figure 7. In an otherwise unrelated experiment in which insertions were made into several muscles of the tongue and pharynx as well as the LCA, subject TB produced some systematic fundamental frequency changes. This figure shows the EMG activity of several muscles--the lateral cricoarytenoid (LCA), levator palatini (LEV), styloglossus (SG), inferior longitudinal of the tongue (IL), mylohyoid (MH), inferior constrictor of the pharynx (IC), superior constrictor (SC), and genioglossus (GG)--during arpeggios in the high extreme of the subject's range. The lineup point is the onset of phonation of the highest note. Although the activity of several muscles is correlated with fundamental frequency, at least some of these (such as the LEV and the intrinsic tongue muscles) are sufficiently unrelated to the larynx that they are unlikely to directly affect F_0. Rather, they seem to reflect a general increase in muscle activity in the head and neck when "reaching" for the highest notes. Although this is an extreme example, it might serve to warn that caution must be observed in the interpretation of EMG results, especially when trying to impute cause-and-effect between the action of a specific muscle and a specific acoustic result.

REFERENCES

Basmajian, J. and G. Stecko. (1962) A new bipolar indwelling electrode for electromyography. J. Appl. Physiol. 17, 849.

Collier, R. (1975) Physiological correlates of intonation patterns. J. Acoust. Soc. Am. 58, 249-255.

Gay, T., H. Hirose, M. Strome, and M. Sawashima. (1972) Electromyography of the intrinsic laryngeal muscles during phonation. Ann. Otol., Rhinol., Laryngol. 81, 401-409.

Hirano, M., J. Ohala, and W. Vennard. (1969) The function of the laryngeal muscles in regulating fundamental frequency and intensity of phonation. J. Speech Hearing Res. 12, 616-628.

Hirano, M., W. Vennard, and J. Ohala. (1970) Regulation of register, pitch, and intensity of voice. Folia Phoniat. 22, 1-20.

Hirose, H. (1971) Electromyography of the articulatory muscles: Current instrumentation and technique. Haskins Laboratories Status Report on Speech Research SR-25/26, 73-86.

Isshiki, N. (1964) Regulatory mechanism of voice intensity variation. J. Speech Hearing Res. 7, 17-29.

Kewley-Port, D. (1973) Computer processing of EMG signals at Haskins Laboratories. Haskins Laboratories Status Report on Speech Research SR-33, 173-183.

Shipp, T. and R. McGlone. (1971) Laryngeal dynamics associated with voice frequency change. J. Speech Hearing Res. 14, 761-768.

Sonninen, A. (1956) The role of the external laryngeal muscles in length adjustment of the vocal cords in singing. Acta Otolaryngol., Suppl. 130.

The Effect of Delayed Auditory Feedback on Phonation: An Electromyographic Study*

M. F. Dorman,[+] F. J. Freeman,[++] and G. J. Borden[+++]

ABSTRACT

Delayed auditory feedback (DAF) alters the temporal pattern of laryngeal and supralaryngeal muscle activity. In some instances, the alterations are manifest simply in terms of prolonged muscle activity, while in other instances, the normal coherent pattern of muscle contraction is fragmented by rapid oscillations in muscle activity. The amplitude of electromyographic activity is also altered by DAF but changes in activity vary considerably between muscles and speakers. The patterns of EMG activity correlated with dysfluencies under DAF appear substantially different from those patterns found in stuttering.

It is well-known that most normal speakers who hear their speech delayed by about 200 msec become dysfluent (Lee, 1951). The dysfluencies, sometimes termed "artificial stutter," are manifest in increased vocal intensity, prolonged vowels and syllable repetition (Fairbanks, 1955). Individuals who stutter, however, become more fluent when speaking under delayed auditory feedback (DAF) (Neelley, 1961). In this paper, which reports a portion of a long-range study of feedback mechanisms used in the control of speech production, we consider two questions: (1) What is the effect of DAF on the laryngeal and supralaryngeal muscle activity of normal speakers? and (2) How does the disruption of electromyographic (EMG) activity under DAF compare with the disruption of EMG activity found during stuttering?

---

*A version of this paper was presented at the 8th International Congress of Phonetic Sciences, Leeds, England, 17-23 August 1975.

[+]Also Herbert H. Lehman College of the City University of New York, and the Graduate School and University Center of the City University of New York.

[++]Also Adelphi University, Garden City, N. Y.

[+++]Also City College, City University of New York, and the Graduate School and University Center of the City University of New York.

Acknowledgment: We are grateful to Drs. Seiji Niimi and Tatsujiro Ushijima of the University of Tokyo, Japan. This research was supported in part by NIDR grant DE10774.

[HASKINS LABORATORIES: Status Report on Speech Research SR-45/46 (1976)]

With respect to the first question, the most striking effect of DAF is a change in the timing of motor activity. Figure 1 shows EMG activity from the genioglossus (GG) during three fluent productions of the phrase "the application of wet mud." Note that the EMG activity precedes each tongue raising event, and that the EMG signals for the three repititions of a given gesture evidence similar patterns of activity. In contrast, Figure 2 shows GG activity during the phrase "the application of wet mud," spoken under DAF. The normal timing of motor commands has been disrupted: there are longer delays between the peaks of EMG activity. Moreover, the patterns of EMG activity for each repetition of a given gesture are rather dissimilar.

A comparison of Figure 1 and 2 suggests that the amplitude of the EMG signal changes under DAF. Muscle activity generally decreases, especially when the speech is most disrupted, as in the first two repetitions of the utterance. It is of interest that the third production of the utterance was the most fluent and the closest in amplitude to the utterance under normal auditory feedback.

The disruption of the normal temporal pattern of muscle activity under DAF is correlated with two prominent aspects of dysfluency: (1) increased vowel duration and (2) syllable repetition. Therefore, we turn now specifically to the EMG correlates of these two phenomena.

Figure 3 shows the EMG correlates of vowel prolongation under DAF. The recordings are from the posterior cricoarytenoid (PCA), vocalis (VOC), and orbicularis oris (OO) muscles during the utterance "wasp sting." Under normal feedback, the VOC, acting in concert with other vocal fold adductors to produce closure for /a/, was active for approximately 200 msec. The PCA was active to open the folds for the voiceless /sp/. The PCA activity was followed 100 msec later by OO activity for /p/ closure. Under DAF, the /p/ closure and the vowels in both "wasp" and "sting" were prolonged. The VOC activity mirrored the vowel prolongation showing--for example, for /a/--100 msec more activity. For the /p/ closure, the OO evidences three peaks of activity over a 200-msec period, in contrast to the single peak of activity over a 100-msec period under normal feedback. Note that the EMG activity under DAF, for the OO, did not evidence a normal, but simply prolonged, pattern of muscle contraction. Rather, the pattern of activity was altered, evidencing rapid oscillations in muscle contraction.

Let me now turn to an example of syllable repetition under DAF. As shown in Figure 4, under normal feedback the superior longitudinal (SL) peaks, for this subject, for the /l/ in "balmy" and the /ð/ in "weather." Under DAF the utterance was rendered as "balmy weathether." The SL did not evidence two "normal" coherent peaks for each repetition of /ð/, but rather the muscle activity was characterized by rapid oscillations.

We turn now to the question of the relationship between the EMG correlates of dysfluency under DAF and the EMG correlates of dysfluency during stuttering. Freeman and her colleagues (e.g., Freeman et al., 1975) have found generally increased EMG activity, especially for the laryngeal muscles, during stuttering. More important, perhaps, is that the normal reciprocity of laryngeal abductor and adductors was found to be disrupted.

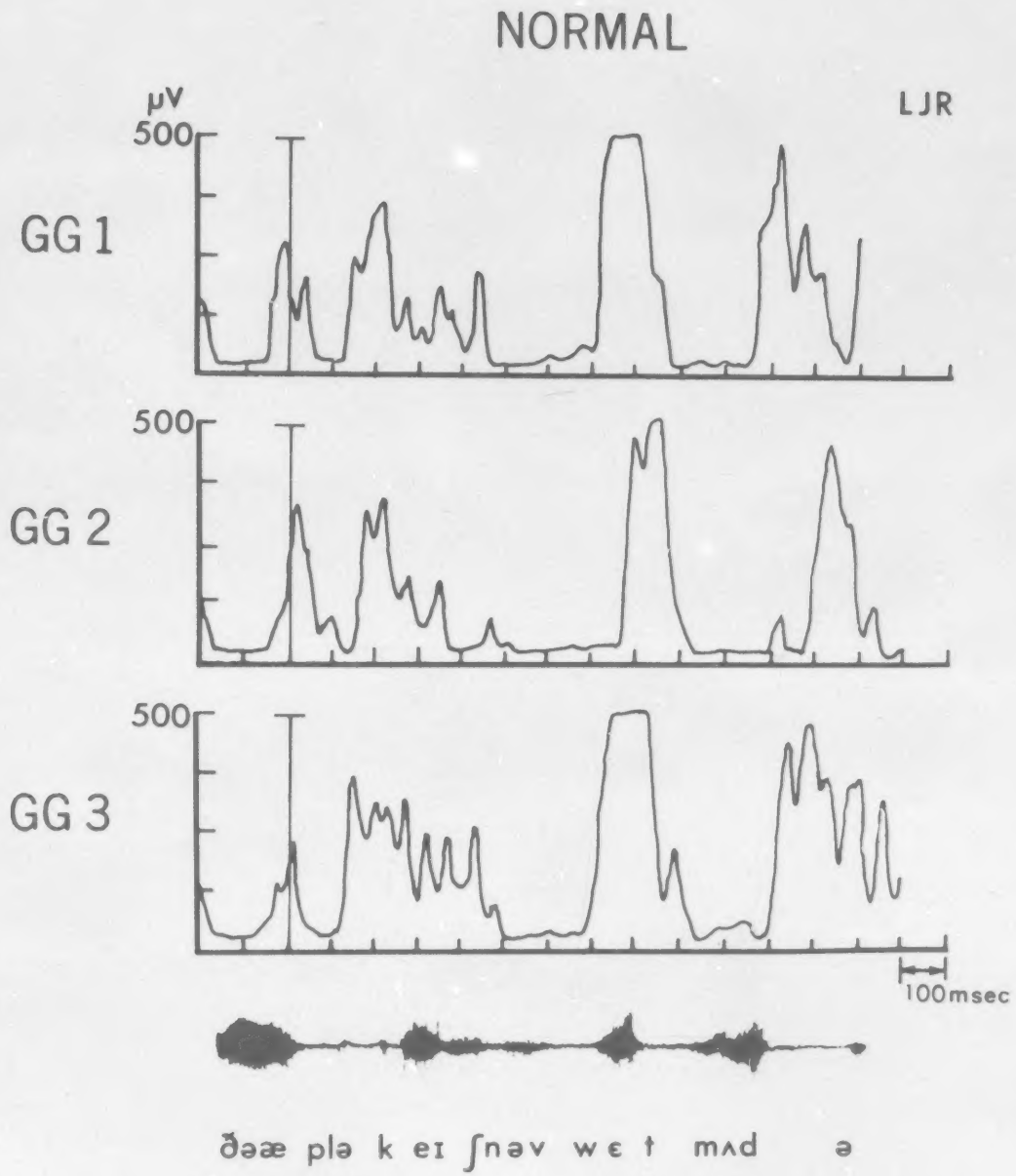


Figure 1: Muscle activity recorded from the genioglossis (GG) during three productions of the utterance "the application of wet mud" under normal auditory feedback.

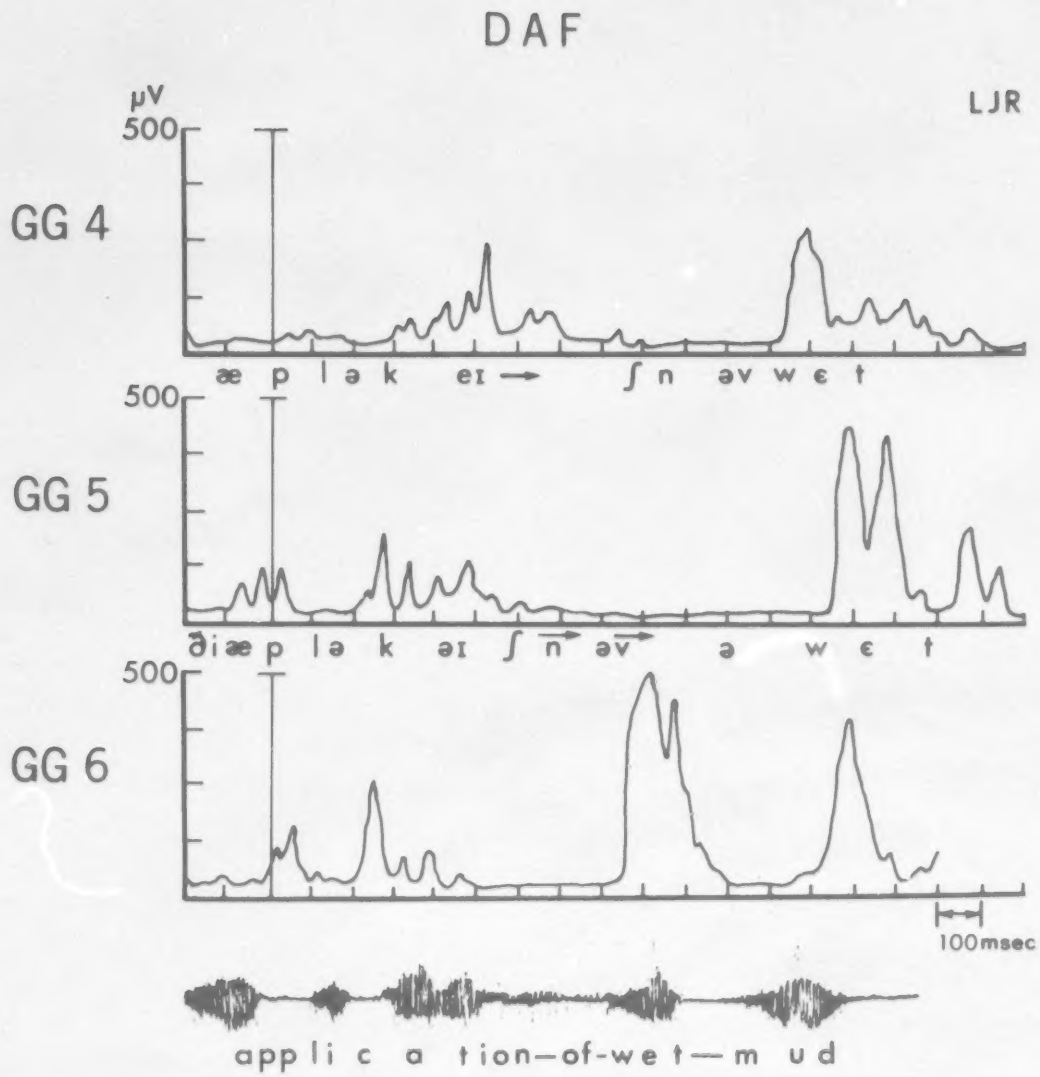


Figure 2: Muscle activity recorded from the genioglossis (GG) during three productions of the utterance "application of wet mud" under DAF.

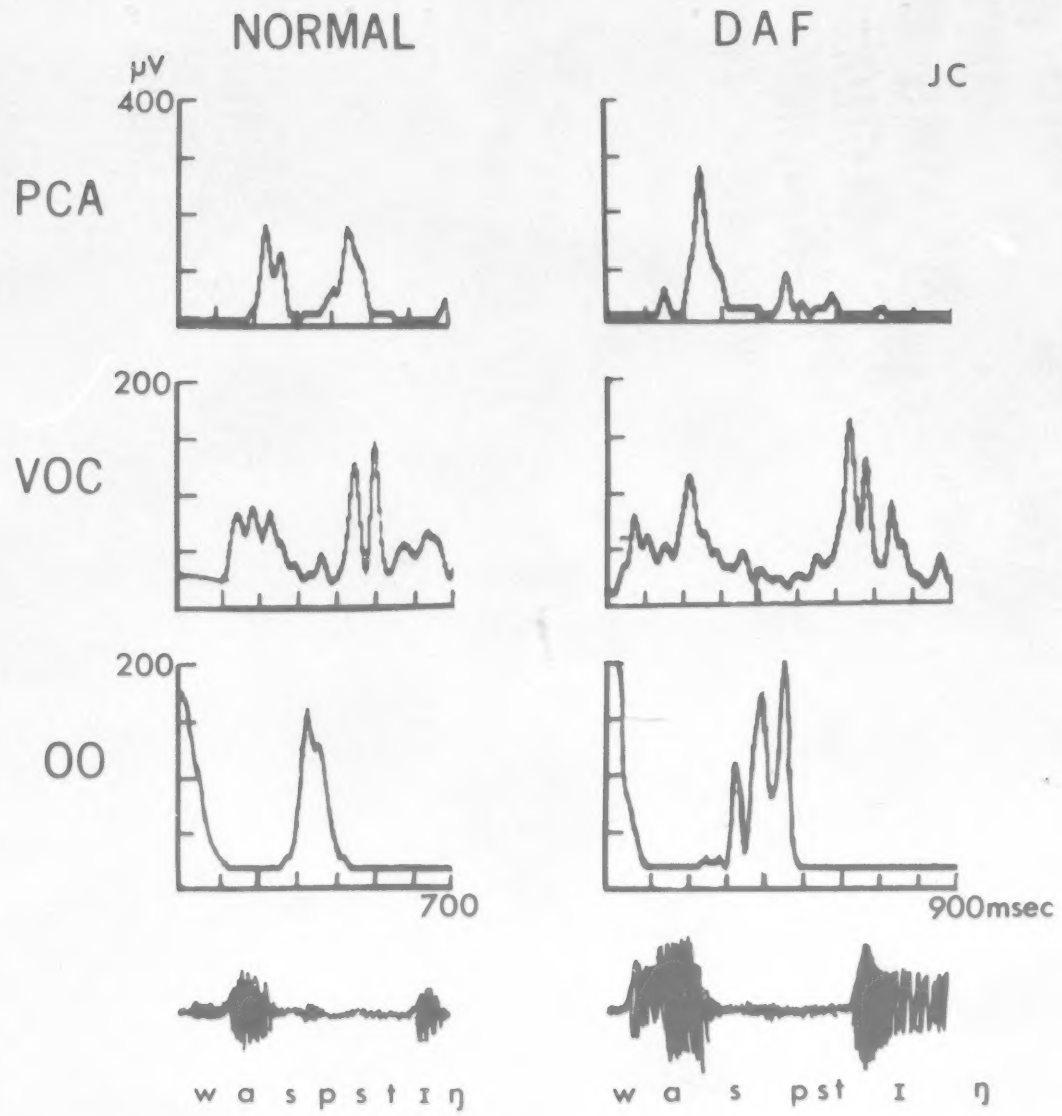


Figure 3: Muscle activity recorded from the posterior cricoarytenoid (PCA), vocalis (VOC), and orbicularis oris (OO) during the production of the utterance "wasp sting" under normal and delayed auditory feedback.

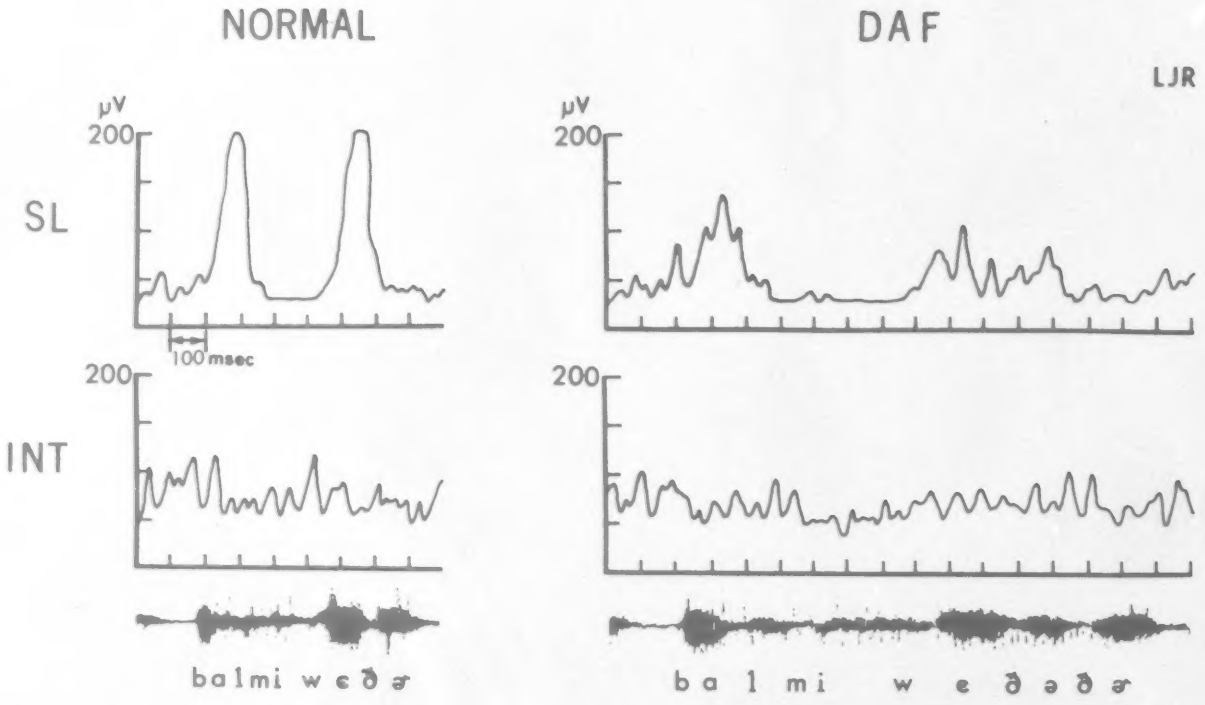


Figure 4: Muscle activity recorded from the superior longitudinal (SL) and interarytenoid (INT) during the production of the utterance "balmy weather" under normal and delayed auditory feedback.

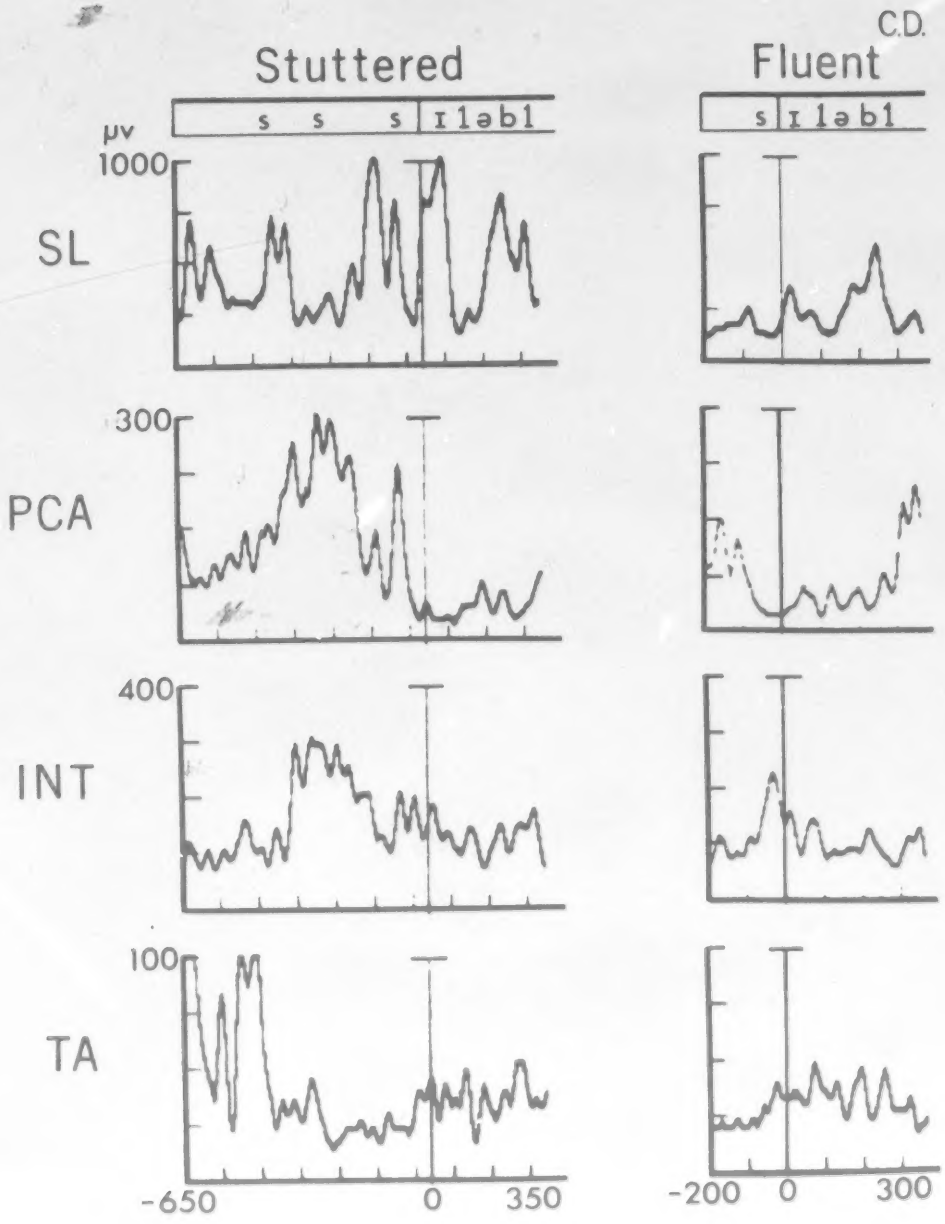


Figure 5: Muscle activity recorded from the tongue (SL), laryngeal adductors (INT and TA), and the laryngeal abductor (PCA) during fluent and stuttered speech.

193

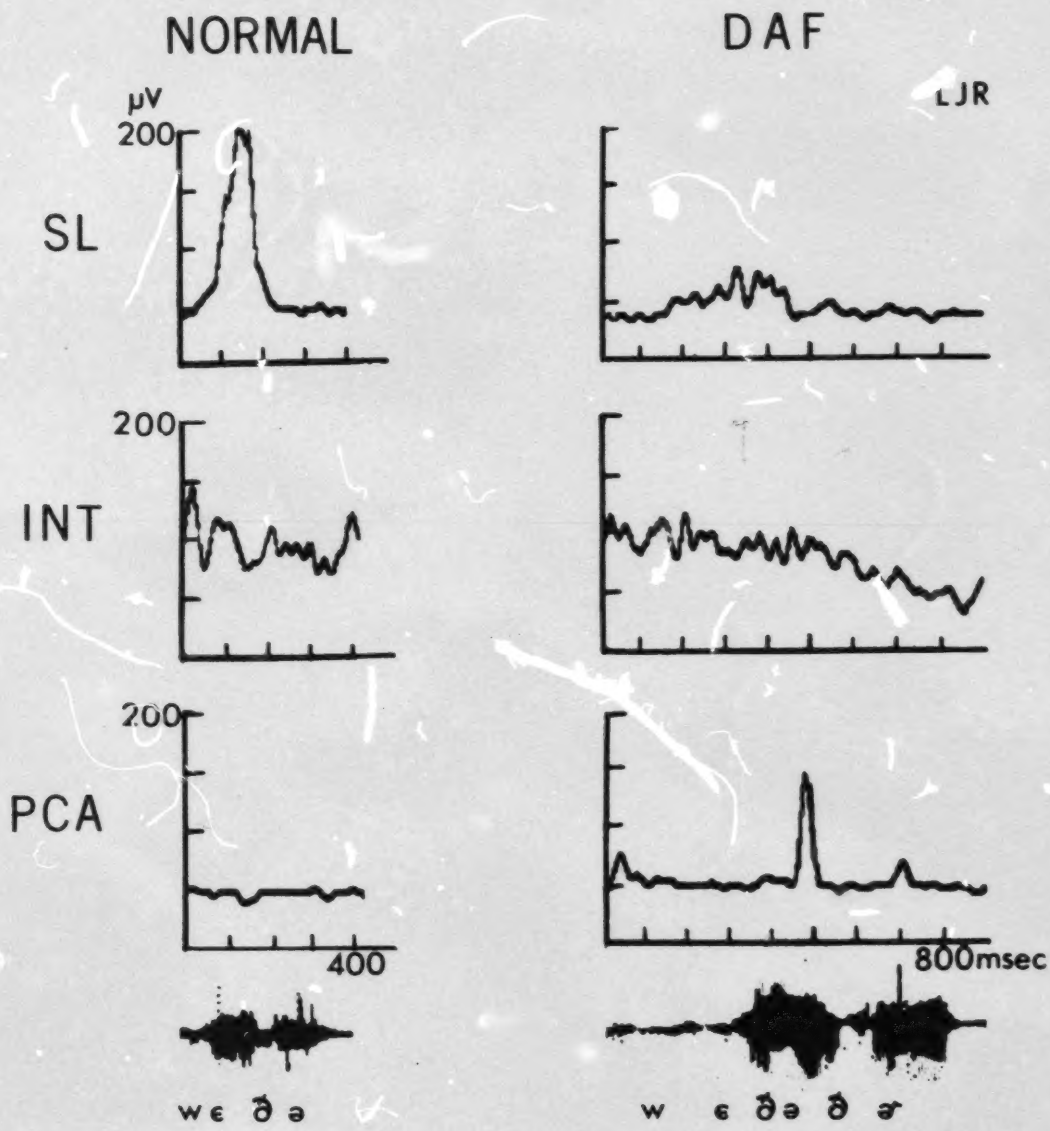


Figure 6: Muscle activity recorded from the tongue (SL), a laryngeal adductor (INT), and the laryngeal abductor (PCA) during the production of "weather" under normal and delayed auditory feedback.

For example, Figure 5 shows EMG recordings from the abductor of the vocal folds, the PCA, and the primary adductor of the vocal folds, the INT. Normally when one is active, the other is inhibited, but during the stuttering block, for example, on the /s/ of the word "syllable," both are active simultaneously. This loss of reciprocity disrupts normal phonation. Amplitude differences between the fluent and stuttered utterances are readily apparent.

Muscle activity, then, for stutterers is generally of higher amplitude during stuttered than during fluent speech, and there is evidence that the normal reciprocal relationship of the abductor and adductor laryngeal muscles is disrupted during stuttering blocks.

The dysfluencies in the speech of normal speakers under DAF are not like stuttering in these two respects. First, under DAF there are amplitude changes in the EMG signal, but the direction of change varies for different subjects and different muscles. For example, Figure 3 indicates an increase in the level of VOC and OO activity under DAF. In Figure 6, however, the SL shows a decrease, the INT shows only minimal changes, and the PCA shows an increase.

The second difference in EMG activity between normal speakers under DAF and stutterers is that during a stuttering block the disruption of reciprocity between abductor-adductor muscles of the larynx prevents or delays normal initiation of voicing while for normally fluent individuals speaking under DAF, voicing usually starts but is either prolonged or "restarted." Typically, in dysfluencies caused by DAF, breakdown of reciprocity occurs after the initiation of voicing. To illustrate, for the fluent production of "weather" shown in Figure 6, the adductor (INT) is active through the utterance because all the segments are voiced. The abductor (PCA) is suppressed throughout the utterance. However, under DAF the abductor fires during the period in which the INT is still strongly active.

To summarize, the main effect of DAF is to alter the temporal pattern of laryngeal and supralaryngeal muscle activity. In some instances the alterations are manifest simply in terms of prolonged muscle activity, while in other instances the normal coherent pattern of muscle contraction is fragmented by rapid oscillations in muscle activity. The amplitude of EMG activity is also altered by DAF but changes in activity vary considerably between muscles and speakers. Finally, the patterns of EMG activity correlated with dysfluencies under DAF appear substantially different from those patterns found in stuttering.

REFERENCES

Fairbanks, G. (1955) Selective vocal effects of delayed auditory feedback. J. Speech Hearing Dis. 20, 333-346.

Freeman, F. J., T. Ushijima, M. F. Dorman, and G. J. Borden. (1975) Dysfluency and phonation: An electromyographic investigation of laryngeal activity accompanying the moment of stuttering. Paper presented at the 8th International Congress of Phonetic Sciences. Leeds, England, 17-23 August.

Lee, B. S. (1951) Artificial stutter. J. Speech Hearing Dis. 16, 53-55.

Neelley, J. N. (1961) A study of the speech behavior of stutterers and nonstutterers under normal and delayed auditory feedback. J. Speech Hearing Dis. Monograph, Suppl. 7, 63-82.

Some Aspects of Coarticulation*

Fredericka Bell-Berti[+] and Katherine S. Harris[++]

ABSTRACT

The analysis of the acoustic and electromyographic experiments reported here indicates that while there is little vowel-to-vowel anticipatory or carryover coarticulation, carryover coarticulation is both more common and more extensive than anticipatory coarticulation.

INTRODUCTION

The nature and extent of coarticulation are of central interest to theories of speech production. Previous work on this problem, for several languages, has shown that anticipatory (or right-to-left) effects extend up to three segments, while carryover (or left-to-right) effects extend up to two segments. In addition, there is evidence that anticipatory effects may be different in cause from carryover effects (for a summary of these data, see Daniloff and Hammarberg, 1973).

More specifically, Kozhevnikov and Chistovich (1965) and Daniloff and Moll (1968) have found anticipatory effects to extend over as many as three phoneme segments and across syllable boundaries. These effects have been explained as the reorganization of motor patterns for speech segments. Carryover effects, on the other hand, have often been attributed to mechanical inertia or articulator "sluggishness" (Lindblom, 1963; Stevens and House, 1963; Henke, 1966; Stevens, House, and Paul, 1966; MacNeilage, 1970), although these effects are now sometimes considered to be deliberate reorganization of speech segments in the same way anticipation is a deliberate reorganization (MacNeilage and deClerk, 1969; Sussman, MacNeilage, and Hanson, 1973; Ushijima and Hirose, 1974).

Despite the central position of coarticulation rules in a general theory of speech production, there are very few descriptive data on the relative magnitude of anticipatory and carryover coarticulation effects at any level. The two experiments presented in this paper provide some of those data. They are extremely similar in the form of the utterances examined. For technical reasons, there

---

*A version of this paper was presented at the 8th International Congress of Phonetic Sciences, Leeds, England, 17-23 August 1975.

[+]Also Montclair State College, Upper Montclair, N. J.

[++]Also The Graduate School and University Center, City University of New York.

[HASKINS LABORATORIES: Status Report on Speech Research SR-45/46 (1976)]

are small differences in the format used in the two experiments. However, as will become apparent, the results for a general theory of coarticulation point in the same direction.

THE ACOUSTIC EXPERIMENT

In the acoustic experiment, the utterance set contained 18 three-syllable nonsense words, consisting of a stressed consonant-vowel-consonant (CVC) preceded by [pə] and followed by [əp]. The vowel in the stressed syllable was either /i/, /ɑ/, or /u/, and the consonants were /p/, /t/, or /k/. All combinations of consonants and vowels were used, except the symmetric ones; for example: /pəpikəp/, /pətupəp/, and /pəkɑtəp/. The utterances were spoken within a carrier phrase, "Say _____ now," at a conversational rate of speech.

Acoustic recordings were obtained, from one speaker of American English, of 18 repetitions of each of the 18 utterance types.

The audio signal was sampled through the Haskins Laboratories pulse-code-modulation (PCM) and Spectrum Analyzing Systems, the former for editing, the latter for generating spectrum data. After software filtering (and thresholding), hard copies of computer-generated spectrograms were obtained and formant measurements made off-line.

Since second-formant (F_2) position is extremely sensitive to back-to-front tongue position and lip-rounding--that is, front cavity length--F_2 measurements were made at seven points in each repetition of each utterance type. Averages of 15-18 measurements for each sample point were obtained. Schematic spectrograms of F_2 were generated from these averages.

The measurement points were

1. One point in $ə_1$ (this syllable was so weakly articulated that no further measures could be made for all utterances);

2. The beginning, middle, and end points of the stressed vowel;

3. The beginning, middle, and end points of $ə_2$.

No attempt was made to account for durational variation, since the sample time represented by each data point in the spectrogram is 12.8 msec; hence, the time scale is too crude for detailed measurements.

RESULTS OF THE ACOUSTIC EXPERIMENT

The results of this experiment are summarized in Figures 1 and 2. Figure 1 shows the 18 utterances plotted with the first consonant held constant; Figure 2 shows the same data with the second consonant held constant. In Figure 1 the left-hand panel represents the averaged F_2 values for utterances whose stressed vowel is preceded by /p/, the middle panel represents the averaged F_2 values for utterances whose stressed vowel is preceded by /t/, and the right-hand panel represents the averaged F_2 values for utterances whose stressed vowel is preceded by /k/. Within each panel, the first schwa is represented by the single points at the left; the stressed vowel in the middle, identified as V' on the abscissa; the second schwa on the right. Second-formant points for

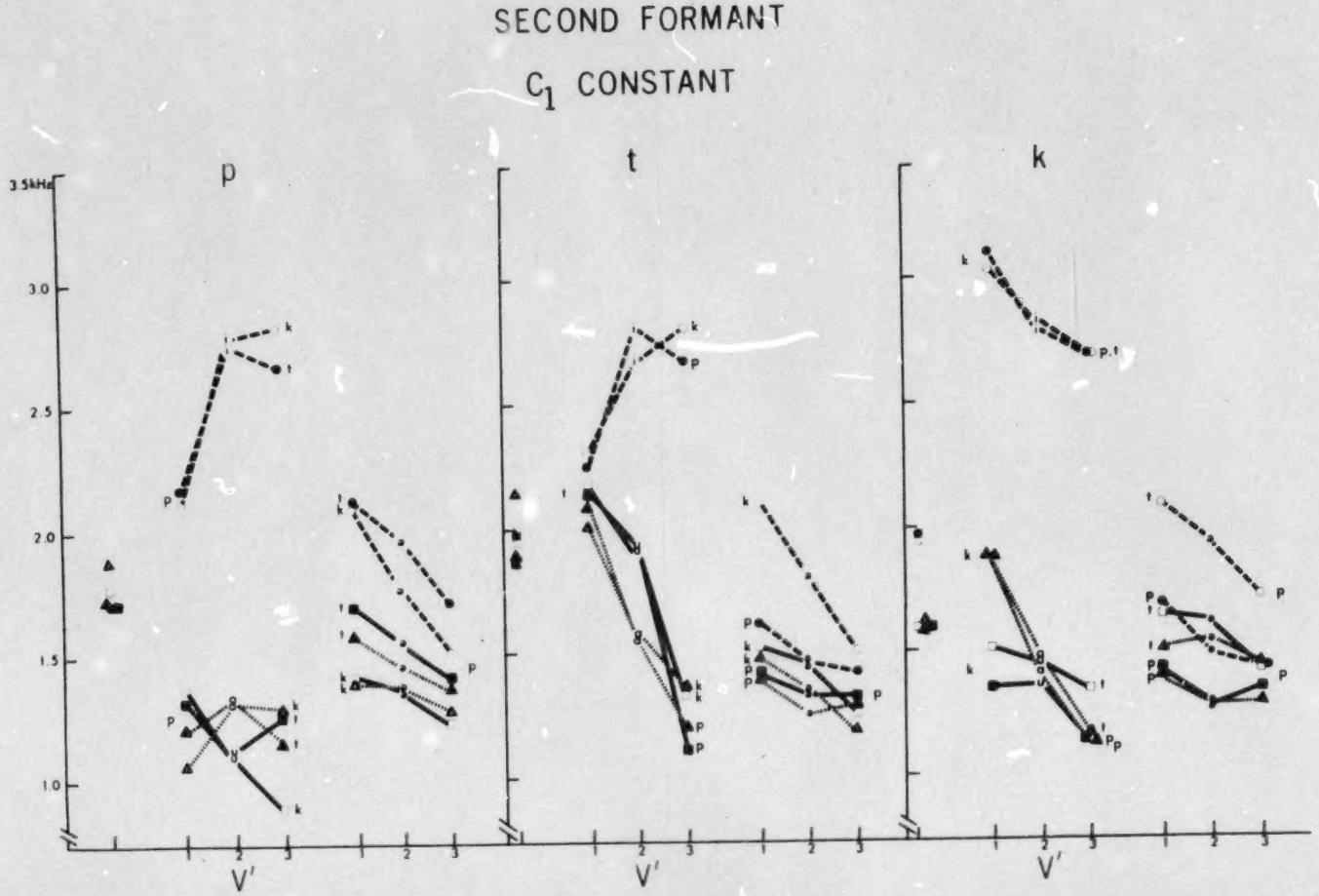


FIGURE 1

Figure 1: All measurements are of F_2. The single data points at the left of each section are $\ominus_1$ measurements. The three-point plots above V' in each section are stressed vowel measurements and the three-point plots to the right of each section are $\ominus_2$ measurements. The left-hand section is F_2 averages for utterances whose stressed syllables begin with /p/; the middle section, with /t/; the right section, with /k/.

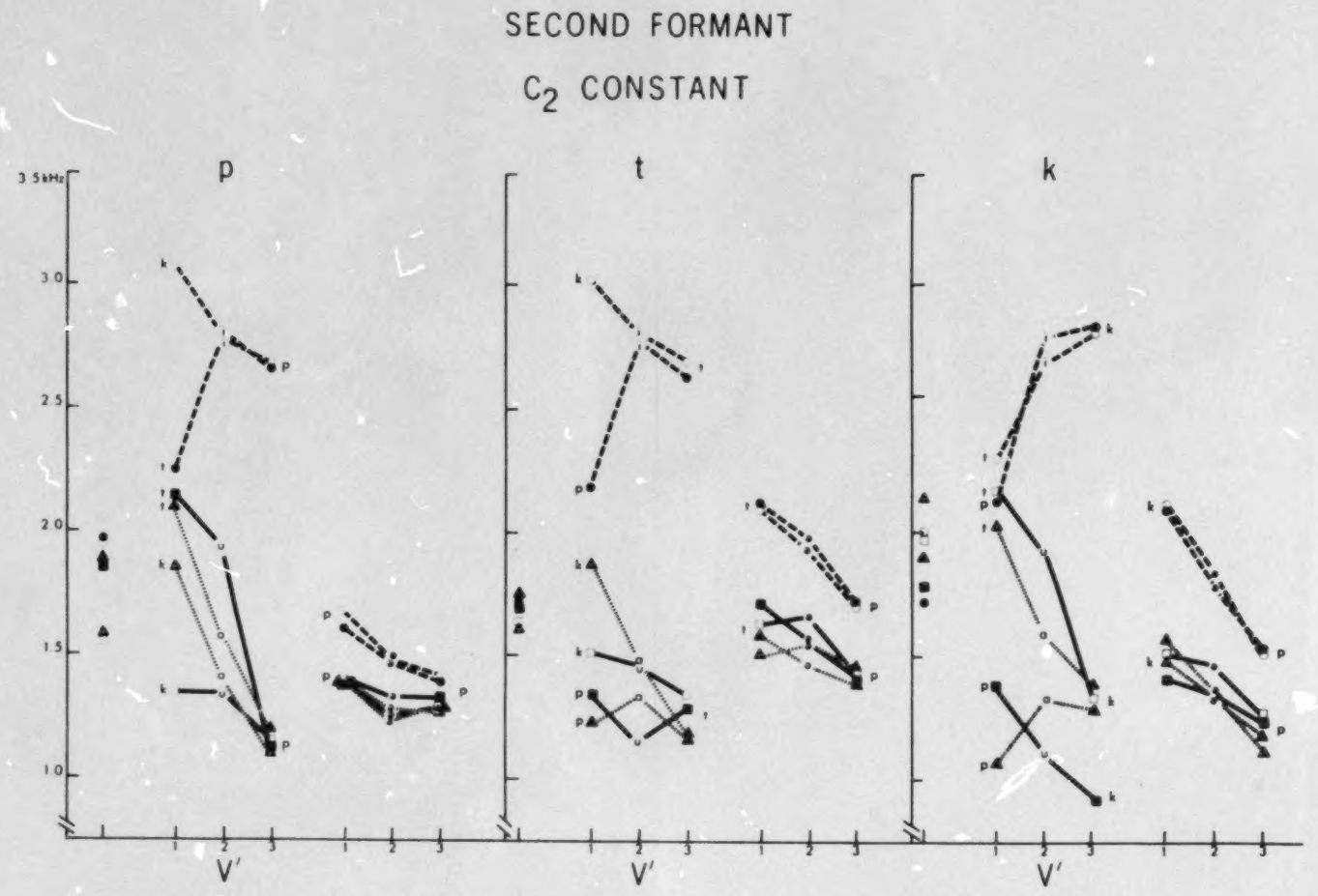


Figure 2: All measurements are of F_2. The single data points at the left of each section are $\ominus_1$ measurements. The three-point plots above V' in each section are stressed vowel measurements and the three-point plots to the right of each section are $\ominus_2$ measurements. The left-hand section is F_2 averages for utterances whose stressed syllables begin with /p/; the middle section, with /t/; the right section, with /k/.

the vowels are marked with circles and connected with dashed lines for utterances whose stressed vowel is /i/, triangles and dotted lines for utterances whose stressed vowel is /ɑ/, squares and solid lines for utterances whose stressed vowel is /u/.

We can examine the relative magnitudes of anticipatory and carryover effects by looking at the effects of the stressed vowel on the initial and terminal schwa vowels. One-step effects are seen in both directions: the initial schwa is affected by the following consonant, while the second schwa is affected by the preceding consonant. However, when we turn to the vowel-to-vowel effects, we find that the initial schwa is not affected by the following vowel: the F_2 averages for $ə_1$ are not separated as a function of the following, stressed, vowel. However, the same stressed vowel does change the value of the following schwa.

In Figure 2 the left-hand panel represents the averaged F_2 values for utterances whose stressed vowel is followed by /p/; the middle panel represents the averaged F_2 values for utterances whose stressed vowel is followed by /t/; and the right-hand panel represents the averaged F_2 values for utterances whose stressed vowel is followed by /k/. Again, within each panel, the first schwa is represented by the single point at the left; the stressed vowel in the middle, identified as V' on the abscissa; the second schwa on the right. Second-formant points for the vowels are connected with dashed lines for utterances whose stressed vowel is /i/, dotted lines for utterances whose stressed vowel is /ɑ/, and solid lines for utterances whose stressed vowel is /u/.

Looking at the second schwa, we find that the second formant is higher, throughout its duration, when it follows /i/ than when it follows /u/ and /ɑ/, regardless of the place of articulation of the intervening consonant.

In general, then, at the acoustic level, carryover effects are larger than anticipatory effects. It is this asymmetry of effect that must be accounted for at the articulatory level.

THE ELECTROMYOGRAPHIC (EMG) EXPERIMENT

The articulatory level we have chosen to examine for manifestations of vowel-to-vowel interaction is the EMG signal. We obtained recordings from the genioglossus muscles of three speakers of American English. The genioglossus muscle, the major muscle mass of the tongue, acts to bunch and raise the tongue, and is most active for high front vowels.

In this experiment there were 24 VCV utterances in which the two vowels were all possible combinations of /i, u, ɑ/ and were always different; the stress was systematically varied between the first and second vowels; the medial consonant was either /p/ or /k/. Additionally, all utterances were preceded by [əp] and followed by [pə], resulting in utterances of the type /əpípupə/ and /əpukápə/.

The data were tabulated by inspecting minimal pairs in which either the first or the second vowel was held constant, and assigning the pairs to the categories: "no difference," "small difference," and "large difference" in EMG activity corresponding to the constant vowel targets of each pair (Figure 3).

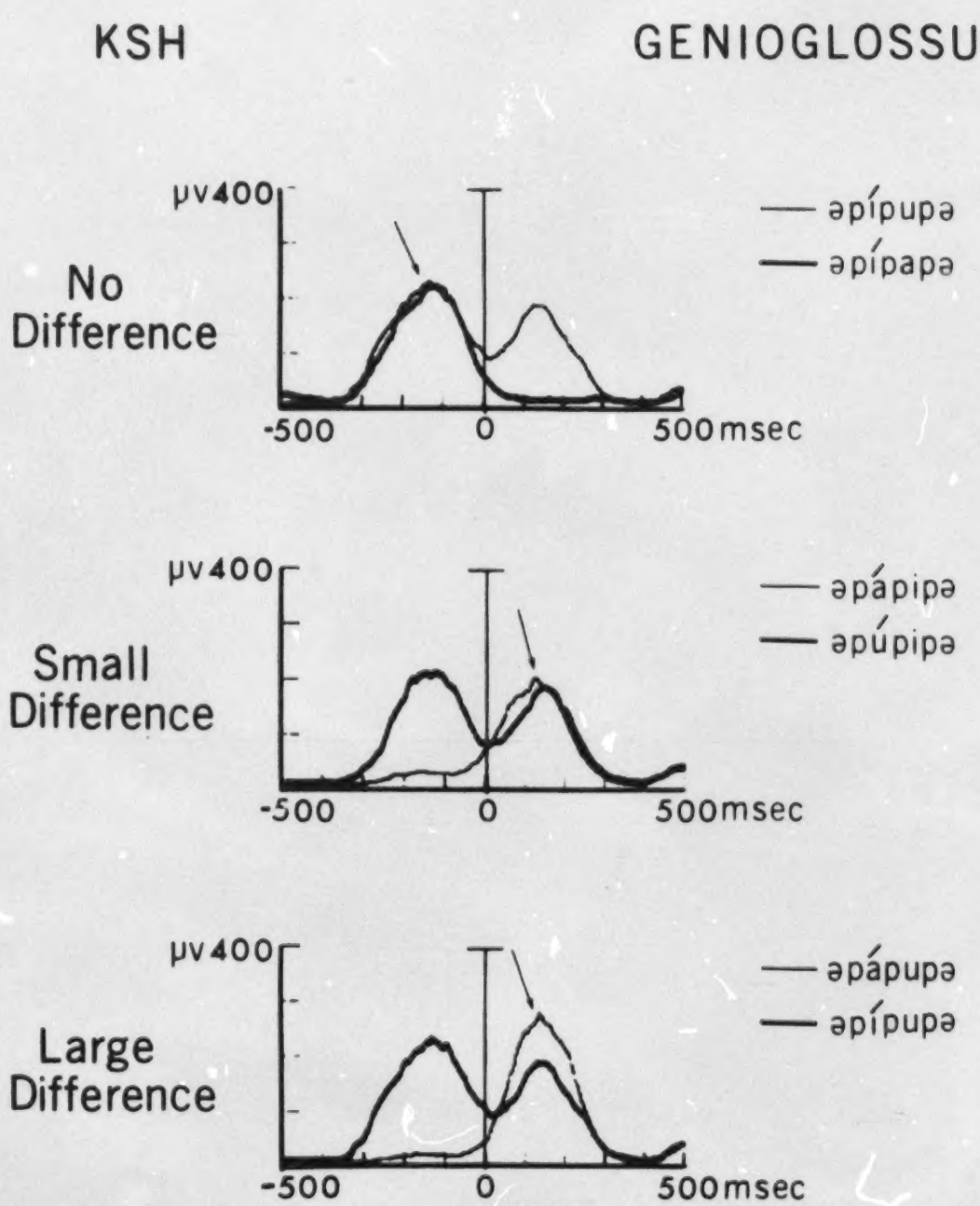


Figure 3: Examples of genioglossus EMG data evaluated as having no difference, a small difference, or a large difference in target vowel activity as a function of quality changes in the nontarget vowel. The top section gives data for anticipatory coarticulation when the target vowel is in the first nonneutral syllable and the middle and bottom sections, for carryover coarticulation when the target vowel is in the second nonneutral syllable.

Both magnitude and timing differences were considered in assigning the contrast pair to one of the categories.

Anticipation was looked for in pairs in which the first vowel was constant; carryover was looked for in pairs in which the second vowel was constant. The number of events in each category was divided by the total number of comparisons to determine the percentage of cases in each of the three categories for both anticipatory and carryover coarticulatory effects.

There was no difference in EMG activity for 75 percent of the anticipatory coarticulation pairs and a small difference in 25 percent of the anticipatory coarticulation pairs (Figure 4). There were no cases in which a large difference in EMG activity was observed in the anticipatory coarticulation pairs. On the other hand, there was no difference in EMG activity for only 25 percent of the carryover coarticulation pairs, a small difference in 45 percent of the carryover coarticulation pairs, and a large difference in 30 percent of the carryover coarticulation pairs.

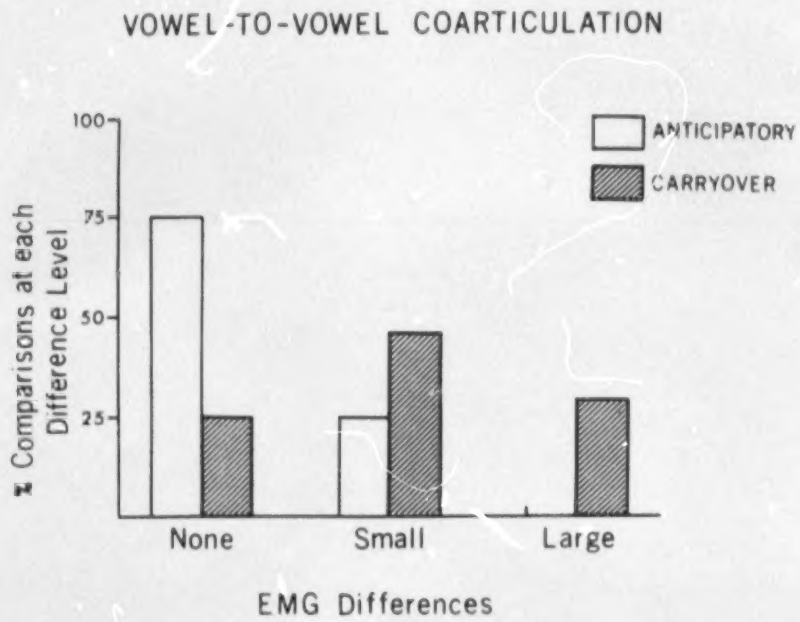


Figure 4: Histogram of proportion of EMG activity magnitude differences for anticipatory and carryover coarticulation.

In other words, there was no vowel-to-vowel anticipatory coarticulation in 75 percent of the anticipatory pairs and there were no large differences in the anticipatory pairs, while there were large differences in 30 percent of the carryover pairs. The results differ somewhat from the EMG data reported by Gay (1975), which may be accounted for by differences in syllable makeup and the rate of speech.

CONCLUSION

Our acoustic and EMG results are in agreement with Gay's (1974) cinefluorographic examination of a very similar corpus, which showed either no vowel-to-vowel coarticulation in either direction, or some carryover coarticulation. These data all support the view that carryover coarticulation is both more common and more extensive than anticipatory coarticulation and is also a reorganization of the motor command.

REFERENCES

Daniloff, R. and R. Hammarberg. (1973) On defining coarticulation. J. Phonetics 2, 239-248.

Daniloff, R. and K. Moll. (1968) Coarticulation of lip rounding. J. Speech Hearing Res. 11, 707-721.

Gay, T. (1974) A cinefluorographic study of vowel production. J. Phonetics 2, 255-266.

Gay, T. (1975) Some electromyographic measures of coarticulation in VCV utterances. Haskins Laboratories Status Report on Speech Research SR-44, 137-145.

Henke, W. (1966) Dynamic articulatory model of speech production using computer simulation. Unpublished Ph.D. dissertation, Massachusetts Institute of Technology.

Kozhevnikov, W. A. and L. A. Chistovich. (1965) (in translation) Speech, Articulation, and Perception. (Washington, D.C.: Joint Publications Research Service, U. S. Department of Commerce, No. 30).

Lindblom, B. E. F. (1963) Spectrographic study of vowel reduction. J. Acoust. Soc. Am. 35, 1773-1781.

MacNeilage, P. F. (1970) The motor control of serial ordering of speech. Psychol. Rev. 77, 182-196.

MacNeilage, P. F. and J. L. deClerk. (1969) On the motor control of coarticulation in CVC monosyllables. J. Acoust. Soc. Am. 45, 1217-1233.

Stevens, K. N. and A. S. House. (1963) Perturbation of vowel articulation by consonantal context: An acoustical study. J. Speech Hearing Res. 6, 111-128.

Stevens, K. N., A. S. House, and A. P. Paul. (1966) Acoustical description of syllabic nuclei: An interpretation in terms of a dynamic model of articulation. J. Acoust. Soc. Am. 40, 123-132.

Sussman, H. M., P. F. MacNeilage, and R. J. Hanson. (1973) Labial and mandibular dynamics during the production of bilabial stop consonants. J. Speech Hearing Res. 16, 397-420.

Ushijima, T. and H. Hirose. (1974) Electromyographic study of the velum during speech. J. Phonetics 2, 315-326.

The Function of Strap Muscles in Speech*

Donna Erickson[+] and James E. Atkinson[++]

ABSTRACT

Association of cricothyroid activity with high or rising fundamental frequency (F_0) and strap activity with low or falling F_0 in speech has been confirmed by numerous electromyographic (EMG) experiments. The purpose of this study is to ascertain whether the role of the strap muscles in lowering F_0 is analogous to that of the cricothyroid in raising F_0. An EMG investigation of the sternohyoid and cricothyroid muscles was performed with speakers of English and Thai. It was found that there were indeed peaks of strap activity during low F_0 and peaks of cricothyroid activity during high F_0. However, examination of the timing of muscle activity with respect to F_0 revealed that the cricothyroid differs from the strap muscles in that the cricothyroid begins to increase in activity prior to the onset of the F_0 rise, whereas the increase of strap muscle activity begins after the onset of the F_0 fall.

It is rather well-known that the cricothyroid muscle is the laryngeal muscle primarily responsible for raising the fundamental frequency (F_0) in speech. There is less agreement as to which laryngeal muscle or muscles is responsible for lowering F_0 in speech. Several electromyographic (EMG) studies with speech have reported an association of strap muscle activity, particularly the sternohyoid, with low F_0, and these studies suggest that the sternohyoid is an active mechanism for lowering F_0. Other studies have shown that there is a decrease of cricothyroid activity associated with low F_0 and have suggested that the cricothyroid is a passive mechanism for lowering F_0.

In this paper we examine more carefully the roles of the sternohyoid and the cricothyroid in lowering F_0. Electromyographic experiments were carried out with a native speaker of Thai and a native speaker of American English. For Thai the utterances examined were the falling tones on three different syllable types, which varied according to vowel and initial consonant: /bîi, pîi, bûu/. Each syllable was preceded by a one-syllable carrier phrase. Figure 1 shows typical

---

*Paper presented at the 90th meeting of the Acoustical Society of America, San Francisco, Calif., 3-7 November 1975.

[+]Also University of Connecticut, Storrs.

[++]Naval Underwater Systems Center, New London, Conn.

[HASKINS LABORATORIES: Status Report on Speech Research SR-45/46 (1976)]

results for the Thai falling tone. It can be seen that there is a decrease in cricothyroid activity and an increase in sternohyoid activity associated with the falling F_0.

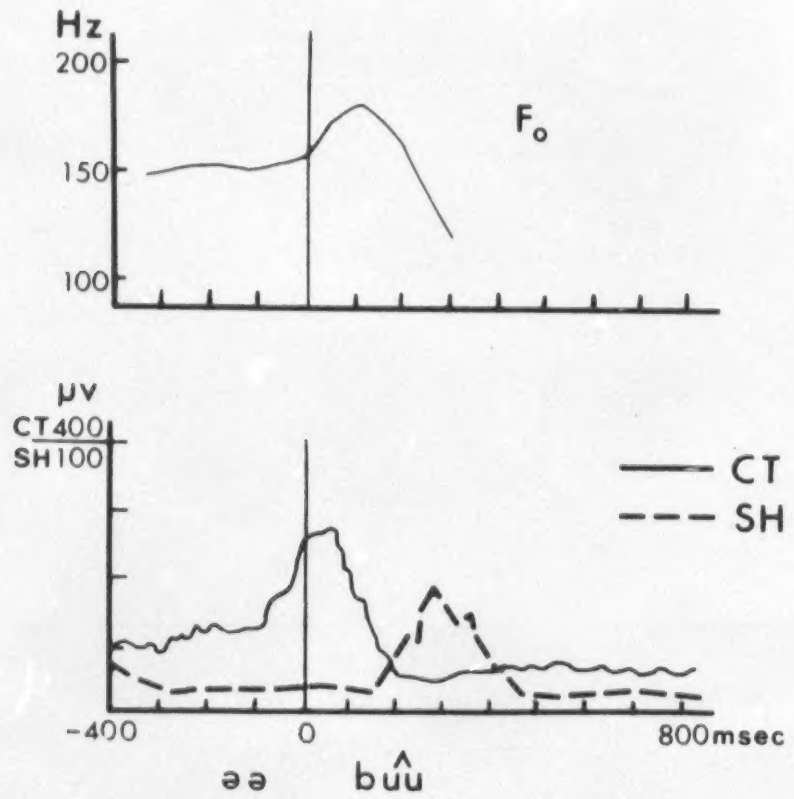
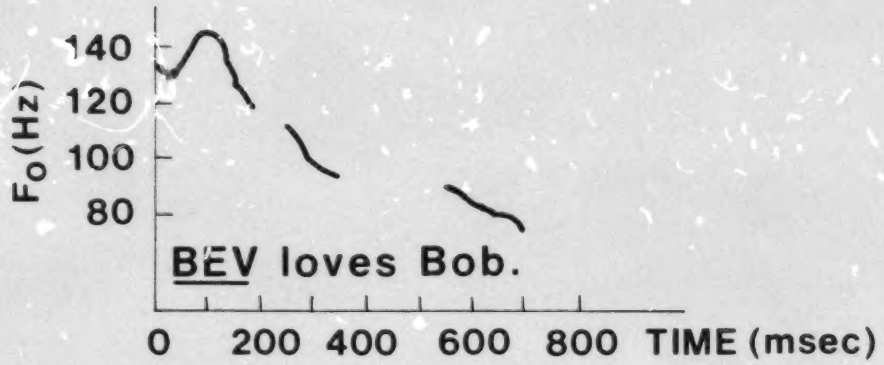


FIGURE 1

Figure 2 shows the utterances examined for English--the falling contours occurred on the stressed words in the sentences "Bev loves Bob" and "Bev loves Bob." Both English and Thai utterance types were chosen from a larger body of data because the onset of the F_0 falls was easily discernible. At least 16 tokens of each utterance type were averaged for English and Thai speakers. Hooked-wire electrodes were used, and the data were processed using the Haskins Laboratories computerized EMG processing system (Hirose, Gay, and Shome, 1971; Kewley-Port, 1973).

In analyzing the data, we looked at the timing of the activity of the sternohyoid and cricothyroid muscles in relation to the F_0 falls. Specifically, as shown in Figure 3, we measured the time at which the cricothyroid activity began to decrease, and the time at which the sternohyoid activity began to increase, both relative to the time at which the F_0 began to fall.


140
120
100
80
F_0(Hz)
BEV loves Bob.
0
200
400
600
800
TIME (msec)

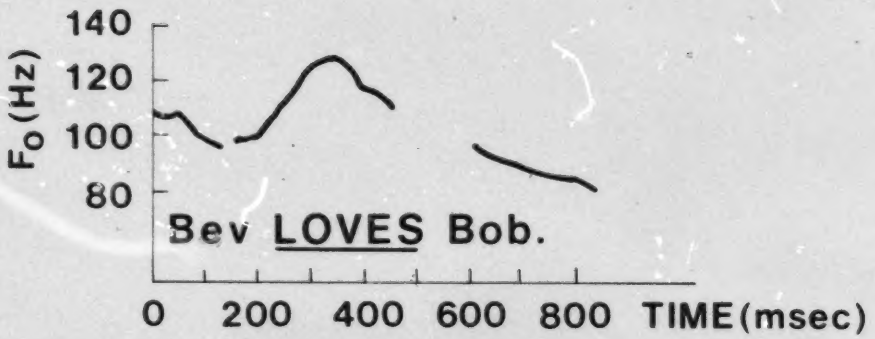
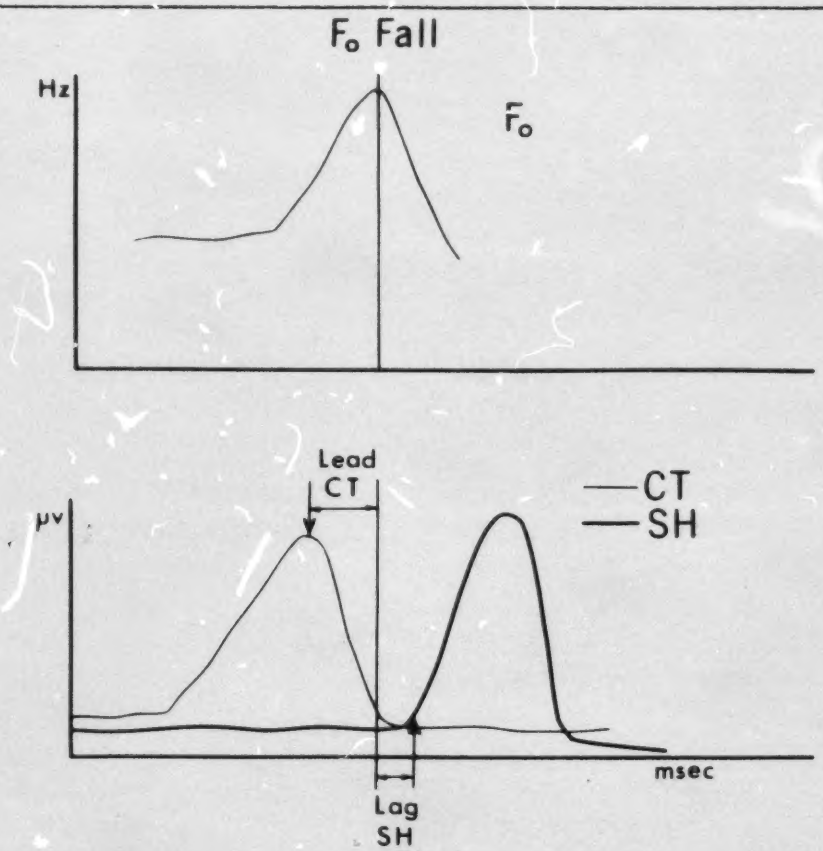

140
120
100
80
F_0(Hz)
Bev LOVES Bob.
0
200
400
600
800
TIME(msec)

FIGURE 2



Schematic presentation of Cricothyroid and Sternohyoid activity in relation to f_o fall.

FIGURE 3

The results for all tokens are shown in Figure 4. The zero reference point indicates the time at which the F_0 begins to fall. It is very important to notice that for both the English and Thai speaker the cricothyroid activity begins to decrease prior to the F_0 fall, whereas the sternohyoid does not begin to increase until after the F_0 fall has begun.

Returning now to our basic question of whether either, neither, or both of these muscles can be responsible for the F_0 fall, it is clear from the above that the cricothyroid begins to decrease before the F_0 fall. It appears, therefore, that the cricothyroid can initiate the F_0 fall by passive relaxation. The sternohyoid, on the other hand, does not begin to increase in activity until after the fall in F_0. Thus it seems that the sternohyoid does not initiate the F_0 falls that we have investigated, although it is clear that the sternohyoid is involved in some way with low F_0.

We feel that we must be careful in interpreting these results not to overgeneralize by implying that the sternohyoid can never initiate falls in F_0. The data in this study are extremely restricted: limited to sharp falls in English in utterance nonfinal position and falling tones in Thai. In both cases

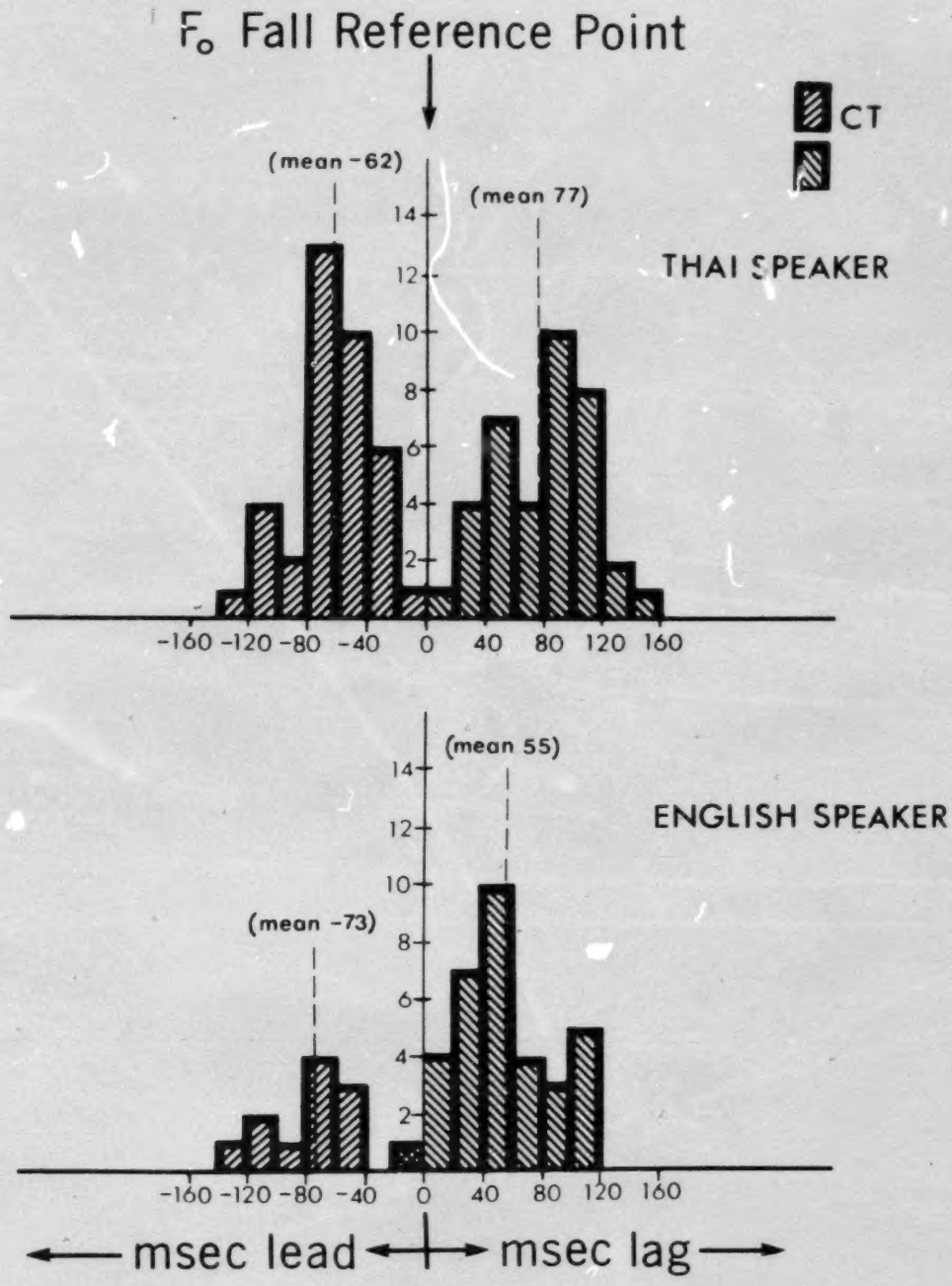

F_o Fall Reference Point
CT
(mean -62)
(mean 77)
THAI SPEAKER
14
12
10
8
6
4
2
-160 -120 -80 -40 0 40 80 120 160
(mean 55)
ENGLISH SPEAKER
(mean -73)
14
12
10
8
6
4
2
-160 -120 -80 -40 0 40 80 120 160
msec lead
msec lag

FIGURE 4

the F_0 falls were from a high to low value. We are expanding the data base to look at what happens when the F_0 falls from a mid to low value. In fact, recent examination of the final fall in the mid tone utterances by the Thai speaker in this paper suggests that there may indeed be instances in which the sternohyoid begins to peak prior to the fall in F_0. This has led us to speculate about a modal shift theory of F_0 lowering. That is, the Thai data suggest that the speaking range can be divided into high, mid, and low voice range, and that an F_0 drop from the high to mid range might be accomplished by relaxing the cricothyroid, whereas a drop from mid to low range involves an increase in sternohyoid activity. This notion will be elaborated in future work.

The mechanism of sternohyoid action in lowering F_0 is not clear. We are still investigating this, as well as other related questions about the strap muscles in speech: Specifically, how do pitch falls interact with jaw opening; how does F_0 interact with vowel and consonant effects; and how do other strap muscles (such as sternothyroid and thyrohyoid) interact with the sternohyoid and each other in these speech activities?

REFERENCES

Hirose, H., T. Gay, and M. Shome. (1971) Electrode insertion techniques for laryngeal electromyography. J. Acoust. Soc. Am. 50, 1449-1450.

Kewley-Port, D. (1973) Computer processing of EMG signals at Haskins Laboratories. Haskins Laboratories Status Report on Speech Research SR-33, 173-183.

Laryngeal Muscle Activity in Stuttering*

Frances J. Freeman[+] and Tatsujiro Ushijima[++]

ABSTRACT

Laryngeal muscle activity during fluent and stuttered utterances was investigated using multichannel electromyography. Analysis revealed that stuttering was accompanied by high levels of laryngeal muscle activity and disruption of the normal reciprocity between abductor and adductor forces. The results demonstrate the existence of a laryngeal component in stuttering and show a strong correlation between abnormal laryngeal muscle activity and perceived moments of stuttering.

INTRODUCTION

For almost a century and a half writers have proposed models of the stuttering block that incorporate an important, perhaps critical, laryngeal component (Arnott, 1828; Müller, 1833; Hunt, 1861; Kenyon, 1943; Moravek and Langova, 1967; Wyke, 1971; and Schwartz, 1974). Recently, an increasing number of studies have indirectly implicated the phonatory mechanism in stuttering (Stromstra, 1965; Wingate, 1969, 1970; Adams and Reis, 1971, 1974; Agnello, 1971; Brenner, Perkins, and Soderberg, 1972).

---

*Two versions of this paper were presented in 1975: "Incoordination and Tension in Stuttering: Further Results of Multichannel Electromyographic Experiments," by F. J. Freeman, G. J. Borden, M. Dorman, S. Niimi, and T. Ushijima, presented at the 50th annual convention of the American Speech and Hearing Association, Washington, D.C., 21-24 November; and "Dysfluency and Phonation: An Electromyographic Investigation of Laryngeal Activity Accompanying the Moment of Stuttering," by F. J. Freeman, T. Ushijima, M. F. Dorman, and G. J. Borden, presented at the 8th International Congress of Phonetic Sciences, Leeds, England, 17-23 August. This article is to appear in The Journal of Speech and Hearing Research.

[+]Also City University of New York and Adelphi University, Garden City, N. Y.

[++]Also University of Tokyo, Japan

Acknowledgment: The authors gratefully acknowledge the invaluable contributions made by Katherine S. Harris and Hajime Hirose. They also wish to thank Norma Rees, Oliver Bloodstein, Irving Hochberg, Gerald McCall, Michael Dorman, Fredericka Bell-Berti, and Diane Kewley-Port for their counsel and assistance.

[HASKINS LABORATORIES: Status Report on Speech Research SR-45/46 (1976)]

Direct evidence of laryngeal involvement in stuttering has emerged from five physiological studies. Chevrie-Muller (1963) used the glottal graph to study 27 stutterers and reported abnormal laryngeal activity that included arhythmic vocal-fold vibrations and unpredictable glottal openings. Fujita (1966) took posterior-anterior laryngeal X rays of a stutterer and found abnormal activity that included irregular and inconsistent opening and closing of the pharyngo-laryngeal cavity and asymmetric tight closure of the glottis. Ushijima, Kamiyama, Hirose, and Niimi (1965); Conture, Brewer, and McCall (1974); and Freeman, Dorman, Ushijima, and Niimi (1975) used the fiberoptic endoscope to view the larynx during stuttering and reported abnormal activity similar to that described by Chevrie-Muller and Fujita. Conture et al. (1974) reported that the abnormal laryngeal activity they observed was suggestive of a disturbance in the smooth, reciprocal interplay between agonist and antagonist laryngeal muscles.

The present research used multichannel electromyography (EMG) to investigate physiological events that occur in conjunction with moments of stuttering. Its primary aim was to describe the laryngeal muscle activity that accompanies stuttering.

METHOD

The EMG techniques used have been developed in a series of experiments investigating normal laryngeal muscle activity in phonation and speech (Faaborg-Anderson, 1957; Hirano and Ohala, 1969; Hirano, Ohala, and Vennard, 1970; Hirose, 1971; Shipp and McGlone, 1971; Gay, Strome, Hirose, and Sawashima, 1972; Hirose and Gay, 1972, 1973). The experimental procedures were described by Hirose (1971) while data collection and processing were discussed by Port (1971) and Kewley-Port (1973, 1974).

Subjects

The subjects for the experiments were four adult males: D.M., P.N., G.G., and C.D. They were selected both because of their willingness to undergo the procedures required for the experiments and because they were anatomically suitable for laryngeal electromyography. The subjects used were the first four suitable individuals located. Subjects G.G. and C.D. were considered mild to moderate stutterers, while D.M. and P.N. were considered severe. They ranged in age from 22 to 47. All had begun to stutter in childhood and each had received some form of therapy.

Procedure

In each case the objective was to secure simultaneous recordings from the five intrinsic laryngeal muscles (cricothyroid, CT; posterior cricoarytenoid, PCA; interarytenoid, INT; thyroarytenoid, TA; and lateral cricoarytenoid, LCA) and at least three of the upper tract articulator muscles (inferior longitudinal, IL; superior longitudinal, SL; genioglossus, GG; and orbicularis oris, OO). Recordings from an extrinsic laryngeal strap muscle, the sternohyoid (SH), were taken for subject G.G.

With one exception (OO for subject G.G.), hooked-wire electrodes (Basmajian and Stecko, 1962) were used. Detailed descriptions of each insertion are given

in Hirose (1971) and Freeman (1975). After each insertion, the electrode-bearing needle was withdrawn leaving the electrodes hooked into the target muscle.

The correct placement of an electrode in a specified muscle was verified in a two-step procedure. First, after each insertion, oscilloscope and amplifier-speaker systems were used for monitoring muscle activity during performance of a series of specified gestures and maneuvers. If the patterns of activity from the insertion site differed from the patterns known to be typical for the target muscle, the electrodes were removed and a new insertion was made for that muscle. Second, recordings were made as the subject performed the critical maneuvers. Using the recordings, final verification was based on examination of the simultaneous activity patterns from each insertion site. Table 1 lists the critical test maneuvers used, and presents a profile of the activity patterns against which each laryngeal insertion was verified. If an insertion could not be verified according to these criteria, the recordings from that site were excluded from the body of data. In cases where spatial proximity makes contamination from adjacent muscles possible, verification was based on demonstrable functional differentiation between the two muscles in question. As indicated by Table 1, functional differentiation is possible between any pair of laryngeal muscles except the LCA and the TA. For these two muscles the patterns are very similar, differing only in degree (level of activity) for some maneuvers.

In addition to the insertion verification procedures, other possible sources of error were considered. Calibration signals of 300 μV, recorded at intervals during the experiments, were compared to verify reliability of recording and playback equipment. The raw EMG tracings were examined visually for (1) abrupt changes in the level of recording from any given muscle and (2) the presence of movement artifacts. Table 2 summarizes the insertions attempted and reports the success rate in achieving verifiable quality recordings from each muscle for each subject.

The design of the study required that comparable fluent and stuttered tokens be obtained from each subject. Since stuttering is a behavior known to be highly variable, the experimental procedures were necessarily flexible.

For subjects P.N., G.G., and D.M. an adequate number of stuttered tokens were obtained by having them read a selected prose passage. Fluent samples were secured by repeated readings (adaptation) and by use of selected fluency-evoking conditions including choral reading, rhythm reading, whispering, reading under white noise masking, and reading under delayed auditory feedback (DAF) (Wingate, 1969, 1970).

Subject C.D. did not have audible blocks while reading the experimental passage. Therefore, he engaged in conversation, making frequent use of feared "difficult" words. In the choral reading condition, the experimenter and C.D. read a list of sentences transcribed from their spontaneous conversation. The recordings made under the other fluency-evoking conditions consisted of spontaneous conversation and repetitions of sentences in which blocks had previously occurred.

RESULTS

The patterns of successful insertion (Table 2) and the procedures used in eliciting fluent and stuttered speech samples yielded results that were not

TABLE 1: Summary of activities used in verification of electrode placement for laryngeal muscle insertions.

MUSCLE	Inspiration	Swallowing	Breath holding	Phonation	Fry phonation	Ascending scale	Descending scale	Jaw opening (resistance)	Head raising (supine)	Speech activities [h a]	[p a]	[b a]	[? a]
PCA	+x	-#	-	-	-	-x	*-			+ -	+ -	- -	- -
INT	-x	+x	+	+	+	++	++			- +	- +	+ +	+ +
LCA	-	+x	+	+	+	++	++			- +	- +	- +	x++
TA	-	+x	+	+	+x	++	++			- +	- +	- +	++
CT	-	-	-		-x	-+x	+-x	-	-				
SH					**x	**x		+x	+x				

+ indicates relatively higher levels of activity
- indicates relatively lower levels of activity or suppression
x indicates a particularly characteristic pattern of activity
# indicates that the maneuver calls for suppression followed by activity
* indicates that at the upper extremes of the subject's singing range activity may occur
** indicates that activity occurs only at the upper and lower extremes of the subject's singing range

TABLE 2: Verified insertions for each subject and for each muscle over the series of experiments.

SUBJECT	Laryngeal muscles						Upper tract articulators				TOTALS*		
	PCA	INT	LCA	TA	CT	SH	IL	SL	GG	OO	Laryngeals	Upper tracts	Combined
D.M.	X		X	X			X	X	X	X	3	4	7
P.N.		X	X	X					X	X	3	2	5
G.G.		X	X	X	X	X		X		X	5	2	7
C.D.	X	X		X				X		X	3	2	5
TOTALS	2	3	3	4	1	1	1	3	2	4	14	10	24

*The Haskins Laboratories multichannel EMG recording and processing system provided for simultaneous processing of recordings from 8 channels. In all cases 8 insertions were attempted. For this series of experiments, 32 insertions were attempted, and of these 24 resulted in successful, verifiable recordings. Two PCA insertions and one INT insertion were impossible because of the subjects' anatomy and gag reflexes; one GG, one LCA, and three CT recordings were rejected because (1) they could not be verified, or (2) they did not result in good quality recordings, or (3) they exhibited evidence of movement artifacts.

parallel for all four subjects. For two subjects (C.D. and D.M.) recordings were obtained for the glottal abductor (PCA) and for glottal adductors. It was possible with these two subjects to study the coordination of the reciprocal activity of the antagonist forces in fluent and stuttered utterances.

With subject C.D., both PCA and INT recordings were obtained for 49 utterances of the same consonant-vowel (CV) sequence allowing a correlation study of abductor-adductor reciprocity in fluent and stuttered utterances.

For the three subjects (D.M., G.G., and P.N.) who stuttered on the oral readings of the experimental passage, it was possible to compare the averaged levels of muscle activity for selected sentences in the stuttered and fluent readings. For C.D. (who did not stutter while reading), the average of the peak values for stuttered and fluent utterances of the same word were compared. These procedures yielded information on two aspects of muscle activity in stuttering: coordination and levels of muscle activity.

Findings Related to Levels of Muscle Activity

In the tracings of the "raw" (unrectified) EMG signal, strength of muscle activity is represented both by the amplitude and frequency of the spikes. Figures 1-3 present examples of raw EMG recordings for D.M., P.N., and G.G. The lower graph in each illustration shows the activity recorded from these same muscles under one of the fluency-evoking conditions. The bottom line in each graph is an oscillographic tracing of the output of the subject's microphone. A phonetic transcription is placed below each graph. In each case the subject was reading the same portion of the experimental passage. Visual inspection of the "raw" EMG data indicates that the laryngeal muscles maintained higher levels of activity during the first (stuttered) reading than during the evoked (fluent) reading.

The differences observed in the "raw" tracings were, of course, apparent in the processed (rectified) EMG. Figure 4 shows recordings from four muscles for subject P.N. The graphs on the left of the illustration traced the course of the EMG activity for those muscles during a stuttered utterance of the word "causes," which occurred in the first (stuttered) reading. The fluent utterance is from his reading under white noise masking. The "raw" EMG for these utterances is shown in Figure 2. Figure 5 shows the activity of a single muscle, the LCA, for three utterances of the word "effect." Subject D.M. repeated the word three times, with progressive adaptation from a severe block to a mild block, to a fluent utterance. The reduction of activity in the LCA correlated with the reduction in degree of dysfluency.

In order to quantify these differences in levels of muscle activity, selected speech samples (consisting in each case of readings of the first paragraph of the experimental passage) were divided into segments of 2-sec duration. The average level of activity in microvolts was calculated for each muscle for each 2-sec segment. The mean values for the 2-sec segments constituting one speech sample were then averaged together, yielding a single mean value for each muscle for each speech sample.

For each speech sample, utterance content was held constant, but the total length of the sample (number of 2-sec segments) varied with utterance rate. For each subject, the first (stuttered) reading was compared with each of the

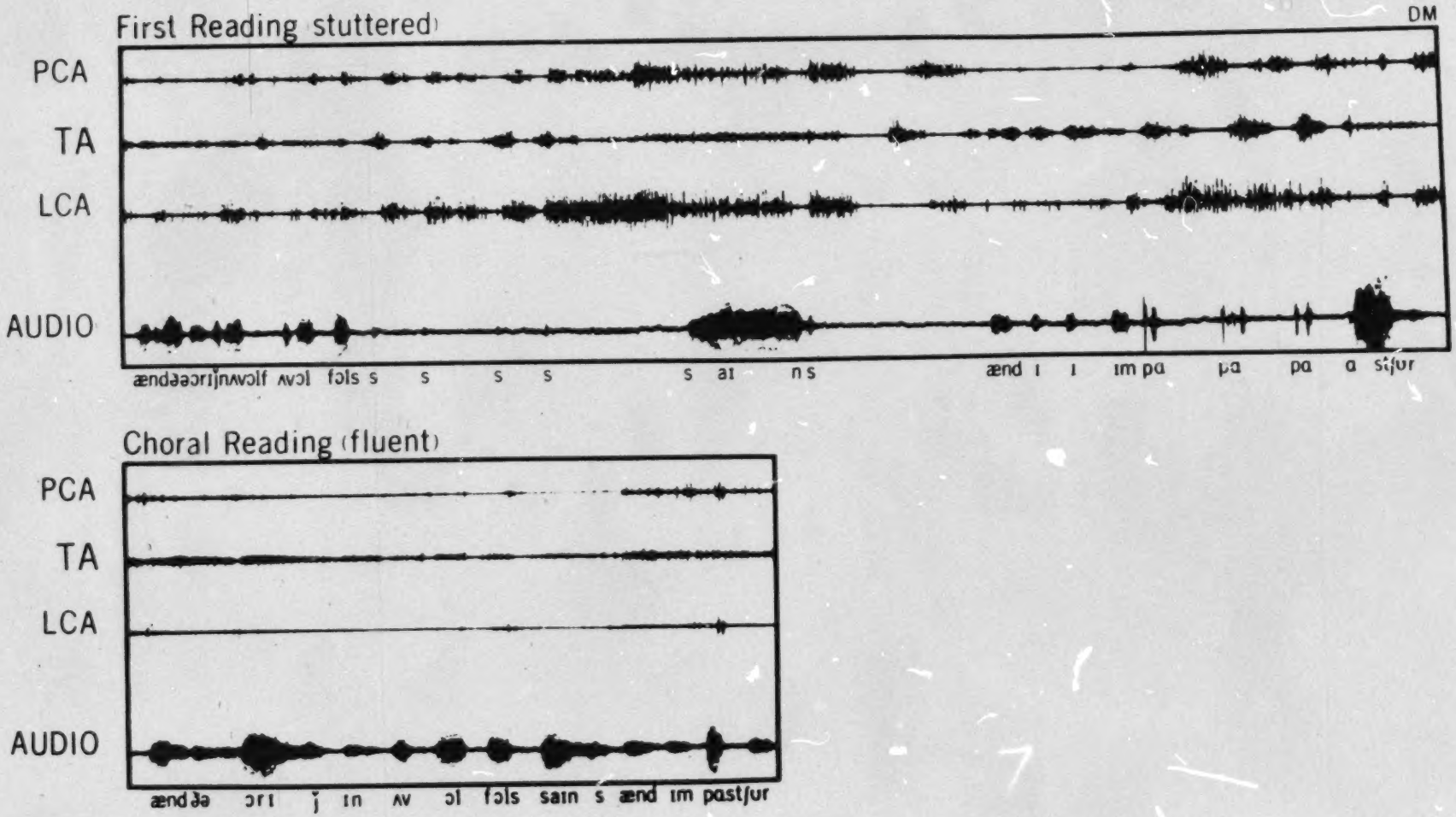


FIGURE 1
214

Figure 1: Comparison of raw (unrectified) EMG recordings from three intrinsic laryngeal muscles--posterior cricoarytenoid (PCA), thyroarytenoid (TA), and lateral cricoarytenoid (LCA)--for subject D.M.'s stuttered and fluent readings of a phrase.

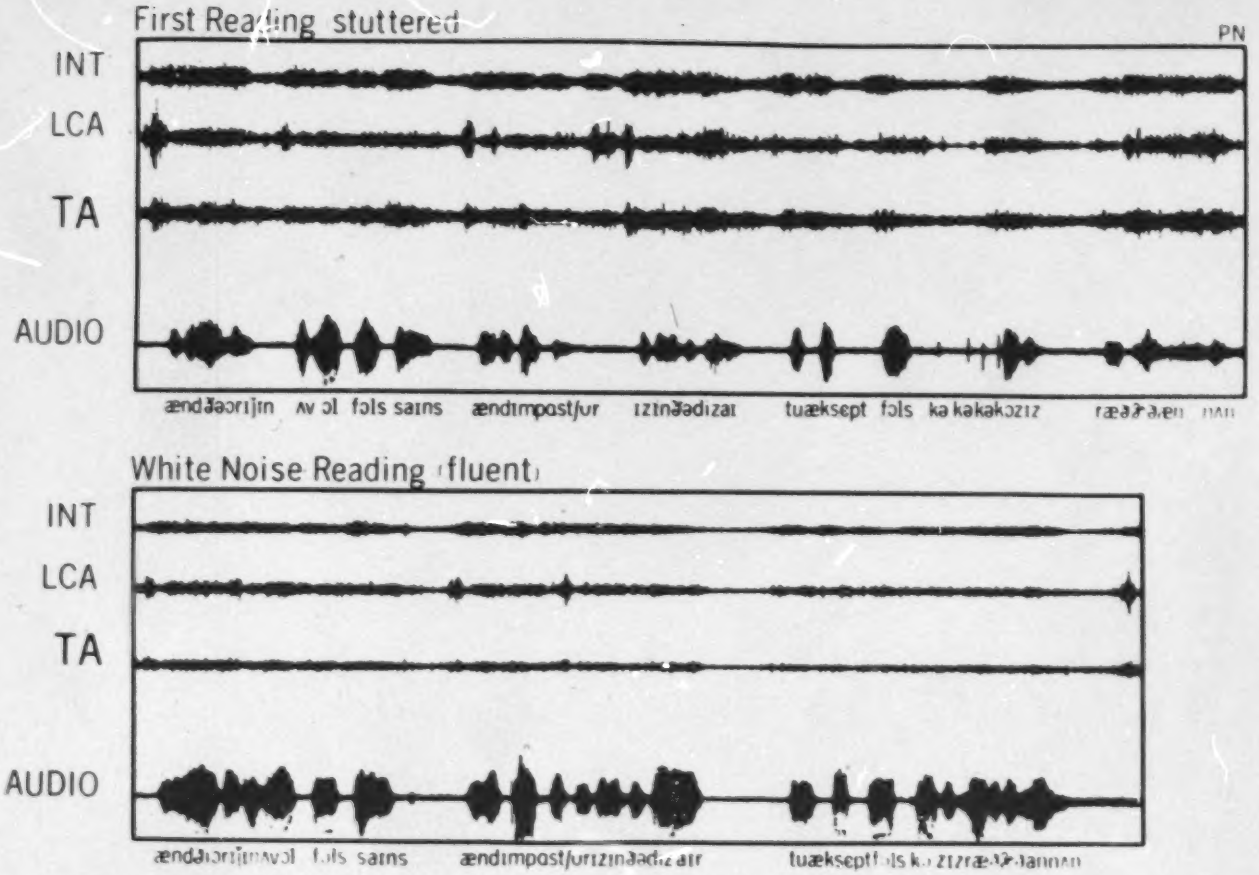


FIGURE 2

Figure 2: Comparison of raw (unrectified) EMG recordings from three intrinsic laryngeal muscles--interarytenoid (INT), lateral cricoarytenoid (LCA), and thyroarytenoid (TA)--for subject P.N.'s stuttered and fluent readings of a phrase.

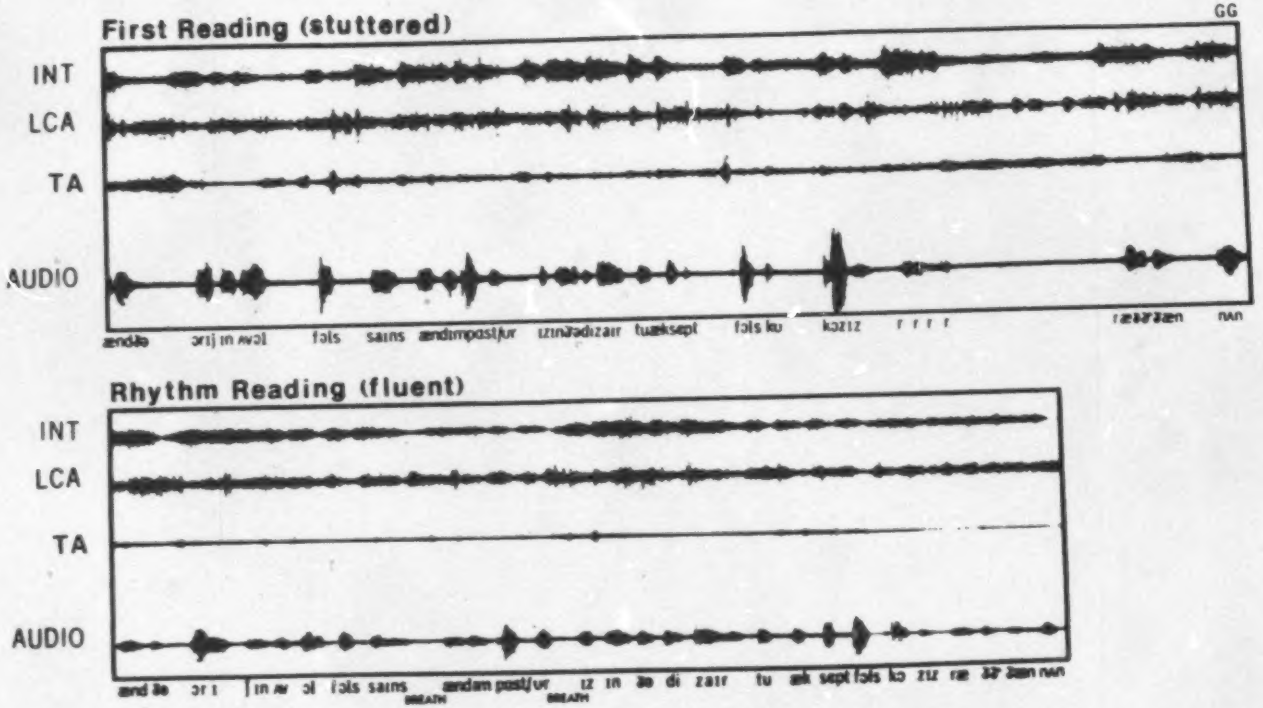


FIGURE 3

Figure 3: Comparison of raw (unrectified) EMG recordings from three intrinsic laryngeal muscles--interarytenoid (INT), lateral cricoarytenoid (LCA), and thyroarytenoid (TA)--for subject G.G.'s stuttered and fluent readings of the same phrase as in Figure 2.

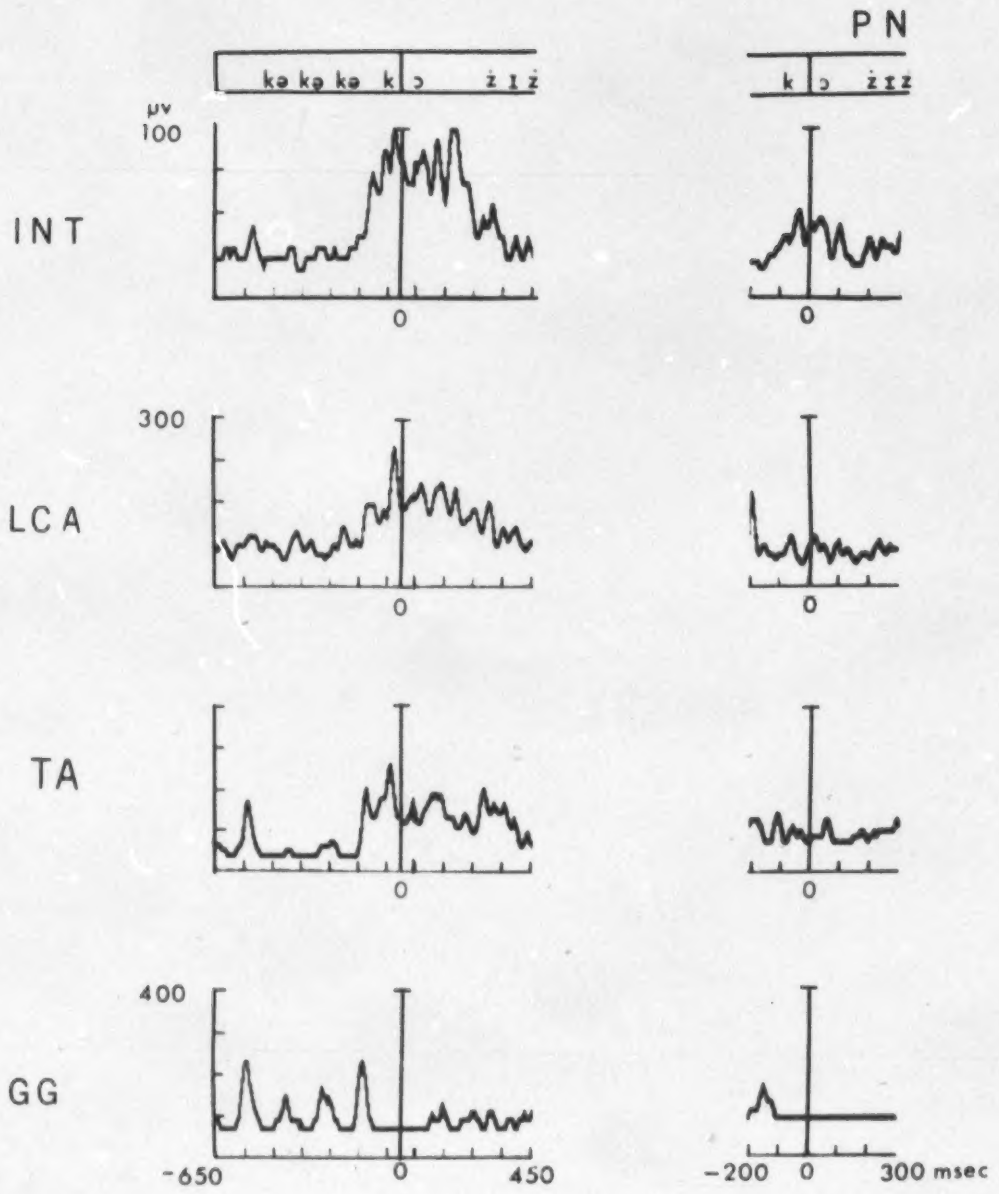


Figure 4: Comparison of muscle activity--interarytenoid (INT), lateral cricoarytenoid (LCA), thyroarytenoid (TA), and genioglossus (GG)--for subject P.N.'s fluent and stuttered utterances of the word "causes."

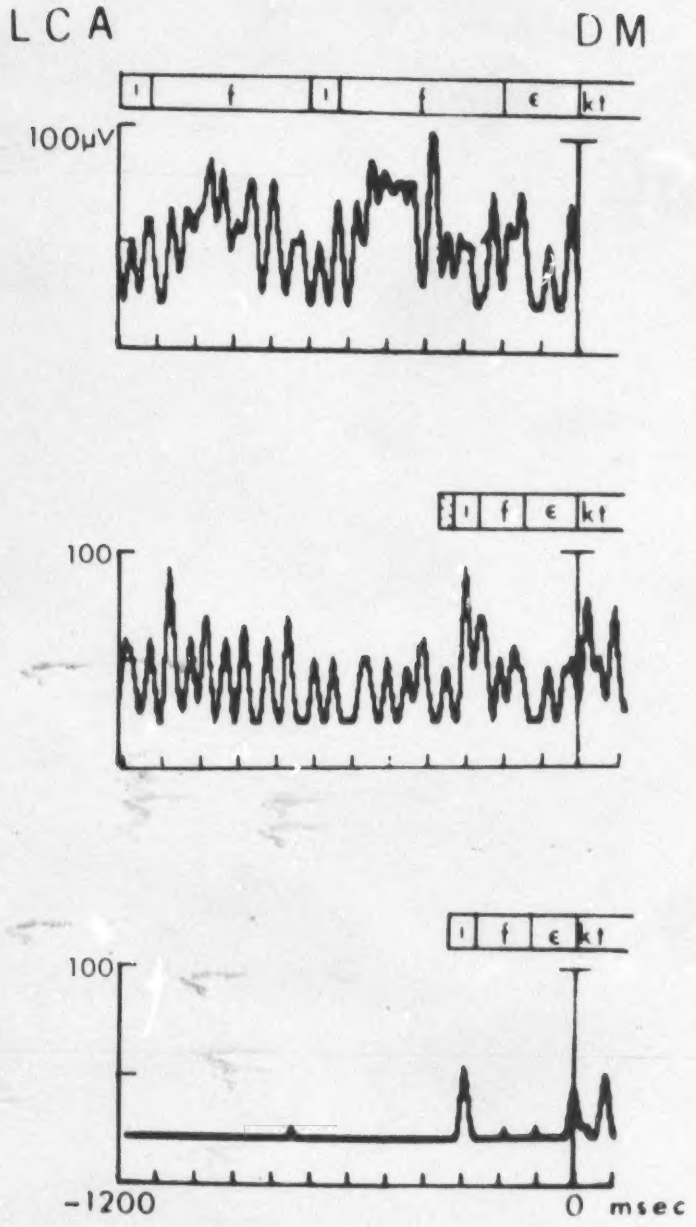


Figure 5: Comparison of lateral cricoarytenoid (LCA) muscle activity for strongly stuttered, mildly stuttered, and fluent utterances of the word "effect" as spoken by subject D.M.

readings under the fluency-evoking conditions. Figure 6 illustrates the differences derived from this comparison, by converting the microvolt values to percentages, using the mean level of the first reading as a reference.

Differences in levels of activity evident in these comparisons are directly related to the two effects of the fluency-evoking conditions on the production of the subjects. In each case, the fluency-evoking conditions resulted in (1) a decrease in the frequency of dysfluencies (measured as percentage of syllables stuttered) and (2) an increase in utterance rate (measured as syllables per second). Figure 7 graphically illustrates these findings for the three subjects. These results, which relate decrease in dysfluencies to increase in rate, are in agreement with a number of other studies of evoked fluency (Adams and Hutchinson, 1974; Conture, 1974). However, the two types of change in the utterance would generate contradictory hypotheses relating to changes in levels of muscle activity. That is, taken alone (without concomitant changes in utterance rate), a marked decrease in stuttering would be anticipated to accompany a decrease in average level of muscle activity. On the other hand, increases in utterance rate will be accompanied by an increase in average level of muscle activity for two reasons. First, an increase in syllables per second results in an increase in the number of speech gestures per second, and hence an increase in the average level of muscle activity per 2-sec segment. Second, an increase in rate results in a higher velocity of articulator movement, which requires a higher level of muscle activity (Bigland and Lippold, 1954; Gay and Hirose, 1973; Kuehn, 1973). Clearly, two opposite and potentially canceling effects were operative simultaneously.

In order to neutralize the effects of the increases in utterance rate, the syllables in each 2-sec segment were counted, and the average level of muscle activity in each segment was divided by the number of syllables uttered in that segment. The resulting means were used to calculate an average level per syllable for each muscle for each speech sample. Results for this calculation are illustrated graphically in Figure 8.

Figure 9 summarizes the results relating to decreases in activity. The broken line labeled 100 percent indicates the reference level of the first (stuttered) reading, while the vertically striated bars are the average of all the upper tract articulator muscles for all the fluency-evoking conditions. The horizontally striated bars are the average of all the laryngeal muscles for all conditions.

The data collected on subject C.D.'s 49 utterances of the word "syllable" and "syllables" were used to learn whether the peak levels of muscle activity were different for fluent and stuttered utterances. In each utterance, the time period between the initial muscle activity for the production of the voiceless fricative [s] (indicated by activity in the superior longitudinal for raising the tongue tip and activity in the PCA for opening the glottis) and the point in the acoustic tract that indicated the onset of voicing for the vowel [ɪ] was identified. Within this time period, the highest peak of activity was identified for each muscle. The level (in microvolts) for this peak of activity was computed for each muscle for each utterance. The experimenter, after listening to audio recordings, identified 23 utterances as stuttered and 26 as fluent. The peak values for the utterances judged stuttered were averaged for each muscle and the results compared with similarly derived averages from the utterances judged fluent. Results are graphically illustrated in Figure 10, where

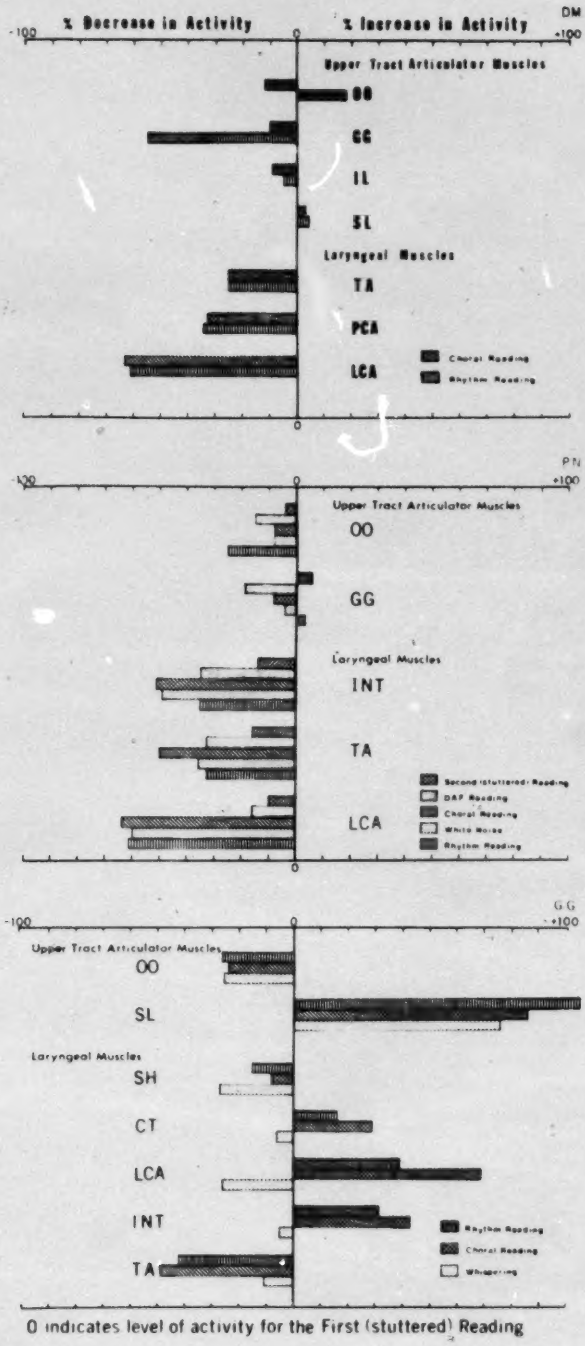


Figure 6: Comparison of average levels of muscle activity per 2-sec segment for subjects D.M., P.N., and G.G.

222

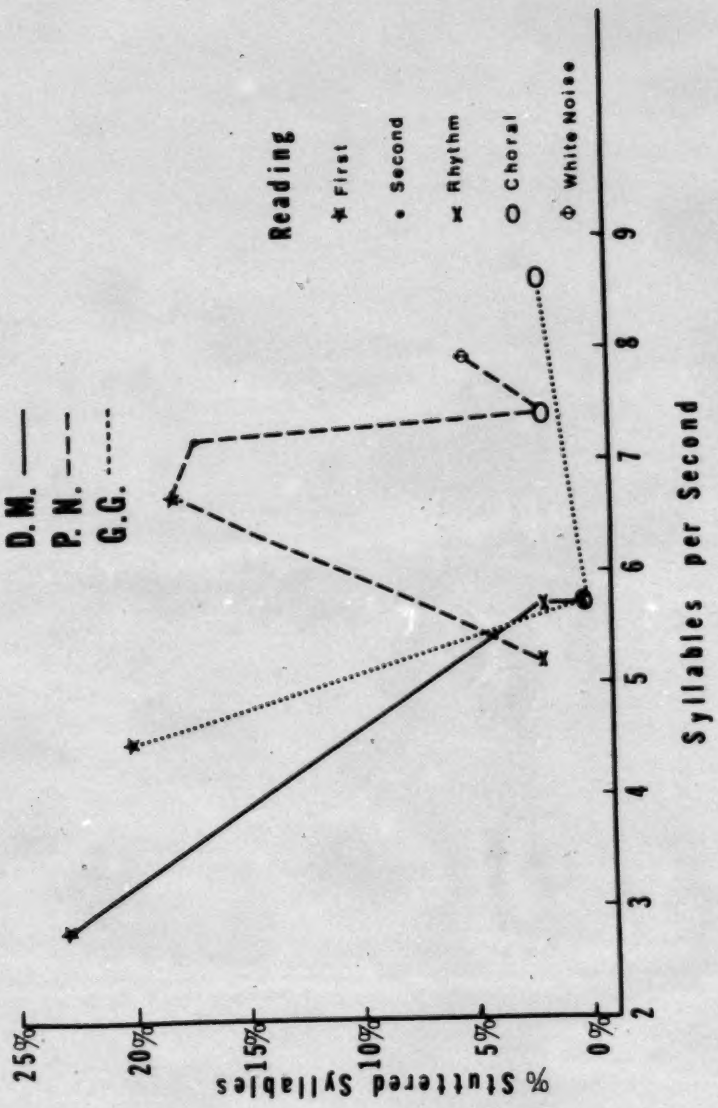


Figure 7: Interaction between utterance rate (measured as syllables per second) and frequency of dysfluencies (measured as percent stuttered syllables).

FIGURE 7

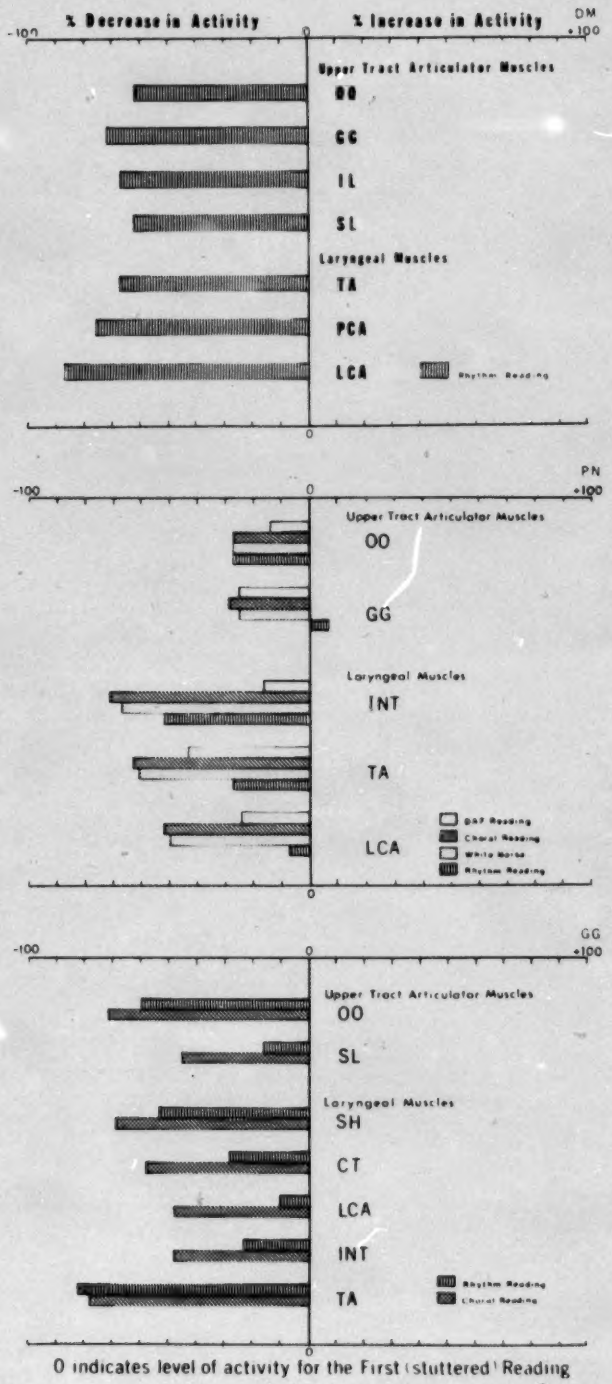
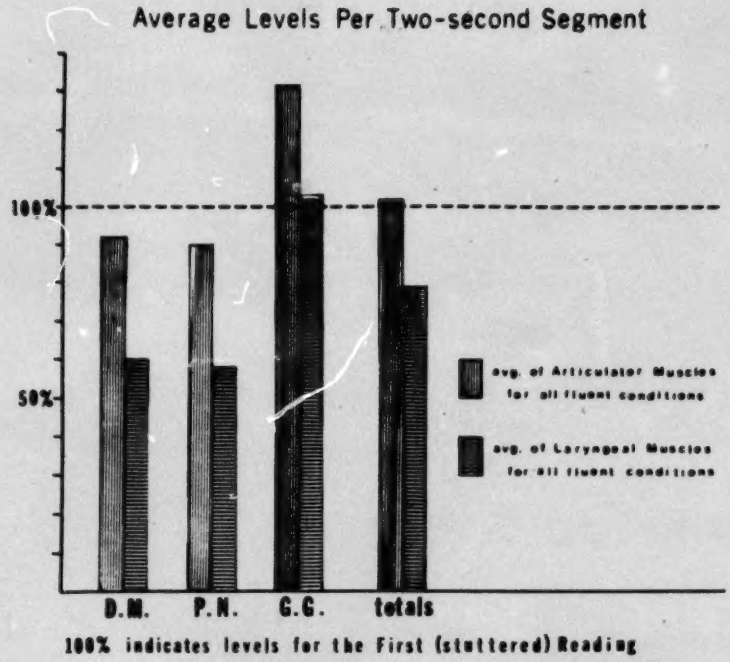


Figure 8: Comparison of average levels of muscle activity per syllable for subjects D.M., P.N., and G.G. [Two reading conditions shown in Figure 6 were omitted here. The second (stuttered) reading was omitted for subject P.N. because it was not significantly different from the first (stuttered) reading; and the choral reading condition was omitted for D.M. because the two voices on the audio recording prevented an accurate syllable count.]



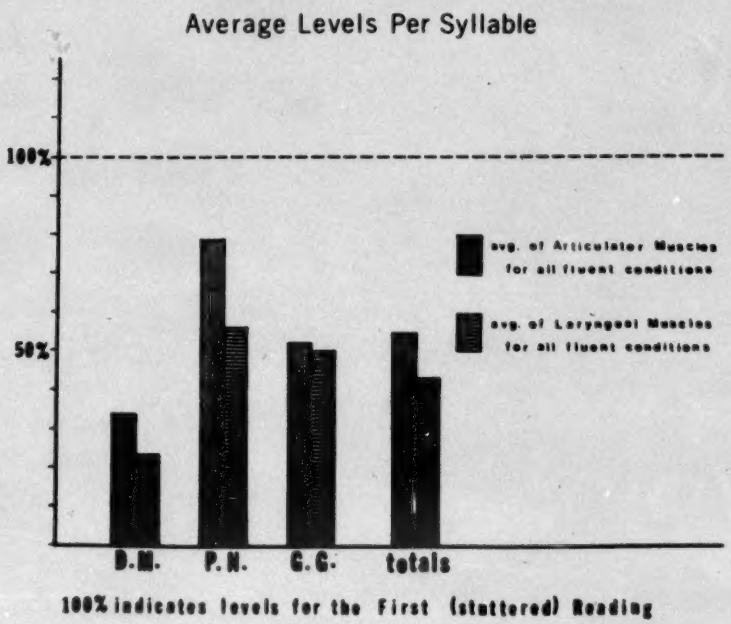


Figure 9: Comparisons of average levels per 2-sec segment and average levels per syllable for upper tract articulator muscles and for laryngeal muscles for subjects D.M., P.N., and G.G.

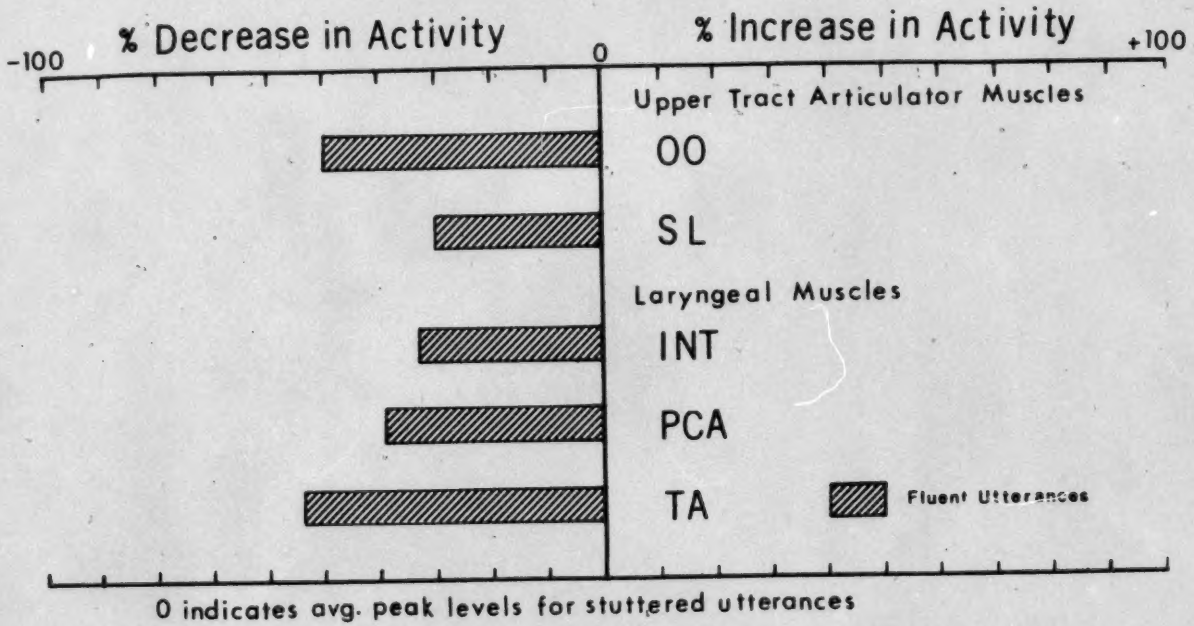


Figure 10: Comparison of average peak levels of muscle activity for subject C.D.'s 23 stuttered and 26 fluent utterances of the words "syllable" and "syllables."

the average peak value for the stuttered utterances serves as the reference and the average peak value for the fluent utterances is expressed as a percent. Differences for four of the five muscles were found to be significant at the .001 level of confidence.

Findings Related to Coordination

The study of disruption of coordination in stuttered speech is restricted to some extent by our imprecise knowledge of many aspects of coordination in normal speech. On one point, however, studies of normal laryngeal articulations have provided relatively clear and consistent findings. These studies indicate that the abductor and adductor forces in the larynx normally act with reciprocity. When the glottal abductor (the PCA) is strongly active, the adductors (INT, TA, and LCA) are suppressed, and conversely, when the adductors are strongly active, the abductor is suppressed. Since recordings from the abductor were secured for two of the four subjects, it was possible to investigate the reciprocal activity of the antagonist muscles.

Figure 11 shows recordings from three muscles for subject D.M. The graph on the left-hand side is from a stuttered utterance of the word, "less," while the graph on the right-hand side is from a fluent utterance of the same word.

The boxes at the top of the graph contain phonetic symbols and represent the relative length of each segment as measured in oscillographic tracings. The lineup, or 0 point, on each graph represents the end of voicing for the vowel. In the bottom graph, the peaks of activity for the SL relate to tongue tip raising for the [l] and the [s]. During the prolongation of the [l] sound, the PCA (glottal abductor) and the TA (a glottal adductor) were both active. During the fluent utterance these two muscles showed reciprocal activity.

Figure 12 shows three utterances of the word "ancient," with progressive adaptation from a strong block to a mild block to a fluent utterance. During the prolongation of the [e] in the strong block, the PCA (glottal abductor) and the TA and the LCA (glottal adductors) were all active. During the fluent utterance the antagonist muscles acted reciprocally.

Figure 13 shows recordings from four muscles for subject C.D. for contrasting stuttered and fluent utterances of the word "syllable." The lineup point for both utterances was on the onset of voicing for the first vowel. In the top graph, the peaks of activity in the SL were related to tongue tip raising. During the stuttered prolongation of the initial voiceless fricative, the PCA (glottal abductor) and the INT (glottal adductor) were both active. During the fluent utterance the antagonist forces acted reciprocally.

The first syllable of the word "syllable" has phonetic content suitable for a correlation study of PCA-INT activity. During the first segment of the syllable, the PCA was active and the INT was suppressed for the production of the voiceless fricative. The INT was then active while the PCA was suppressed for the production of the vowel. This pattern is shown in the fluent utterance of Figure 13. If the normal activity of these antagonist muscles were to be correlated over time, a negative correlation should result. And, indeed, the plotting of such a correlation for the fluent utterance in Figure 13 yielded an r of -.83. Conversely, the plotting of the correlation between the INT and the PCA for the stuttered utterance in Figure 13 yielded an r of +.80.

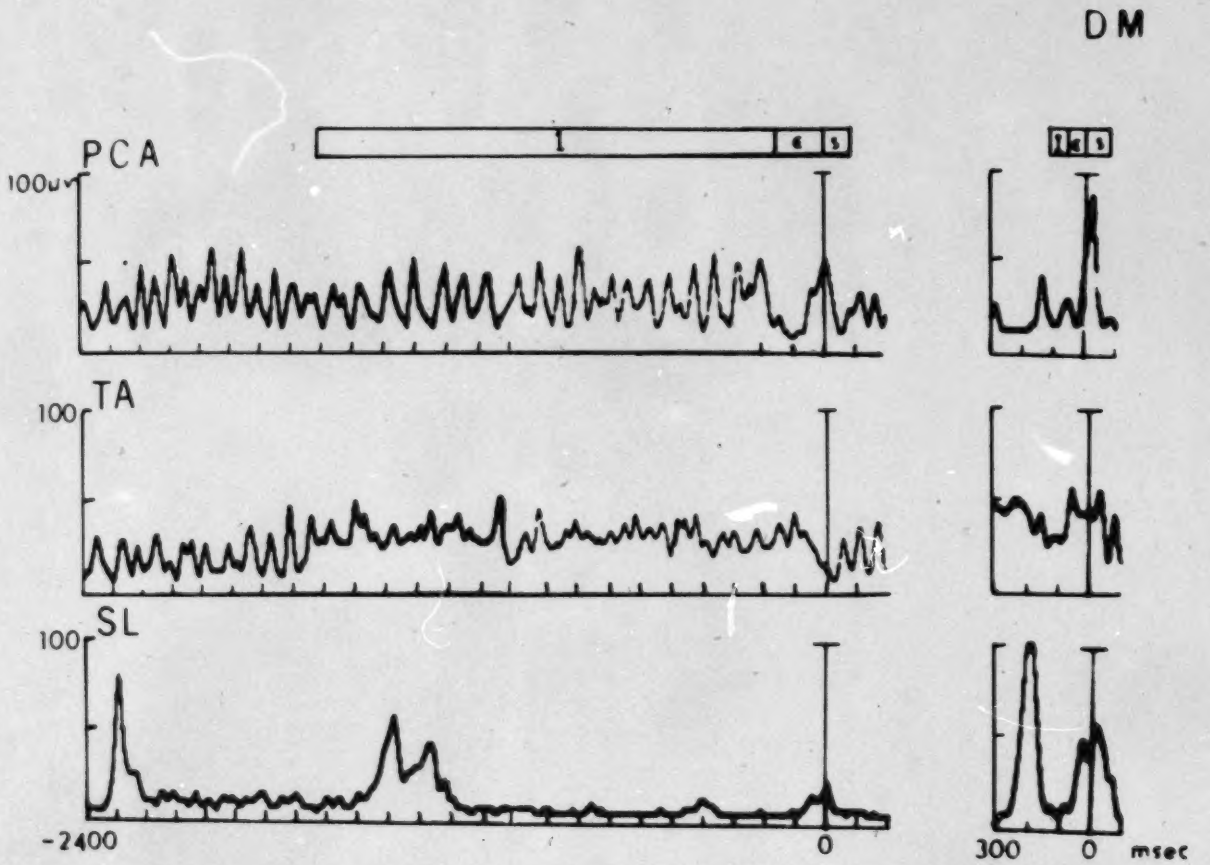


Figure 11: Comparison of muscle activity--posterior cricoarytenoid (PCA), thyroarytenoid (TA), and superior longitudinal (SL)--for subject D.M.'s stuttered and fluent utterances of the word "less."

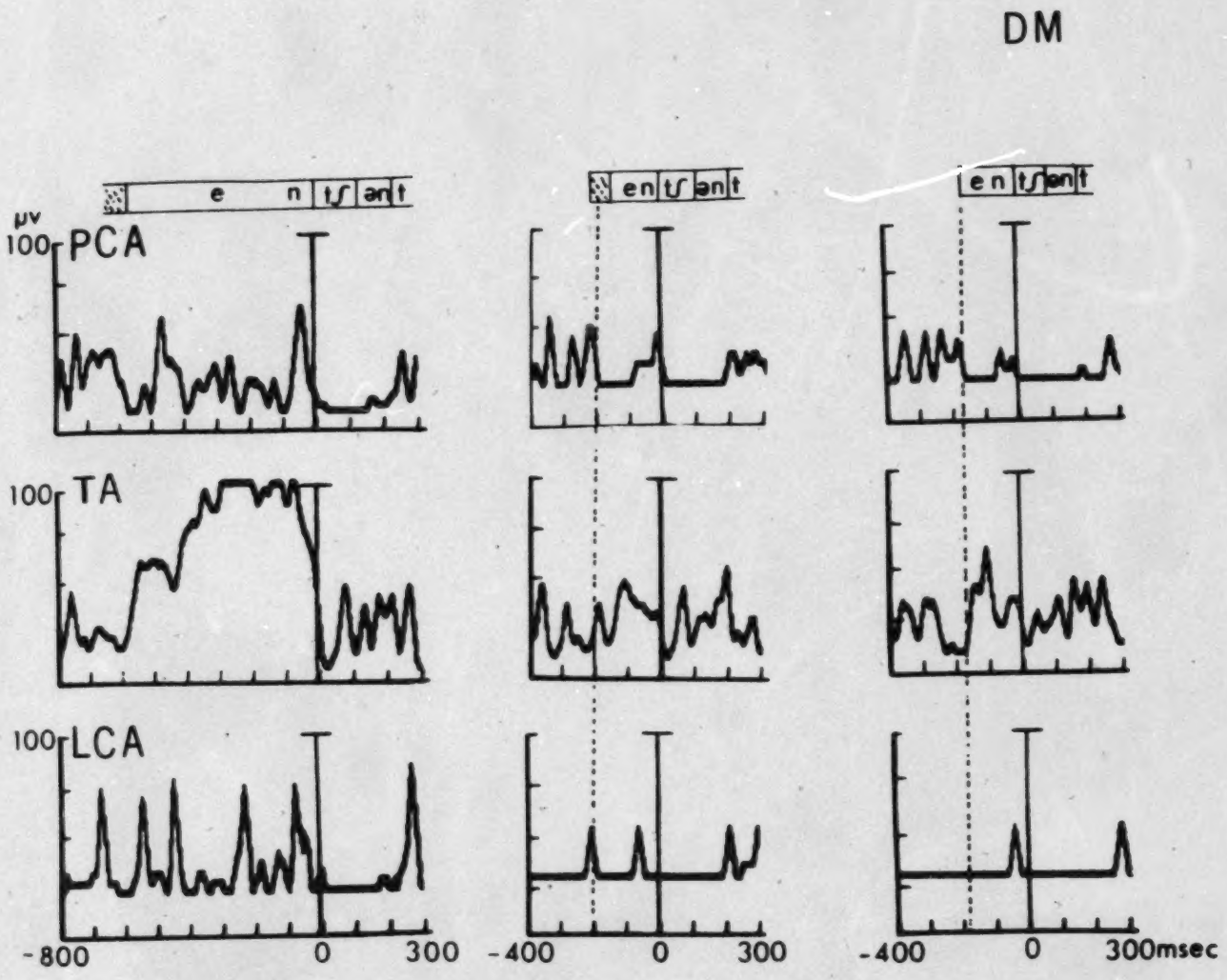


FIGURE 12

227

Figure 12: Comparison of muscle activity--posterior cricoarytenoid (PCA), thyroarytenoid (TA), and lateral cricoarytenoid (LCA)--for strongly stuttered, mildly stuttered, and fluent utterances of the word "ancient" as spoken by subject D.M.

229

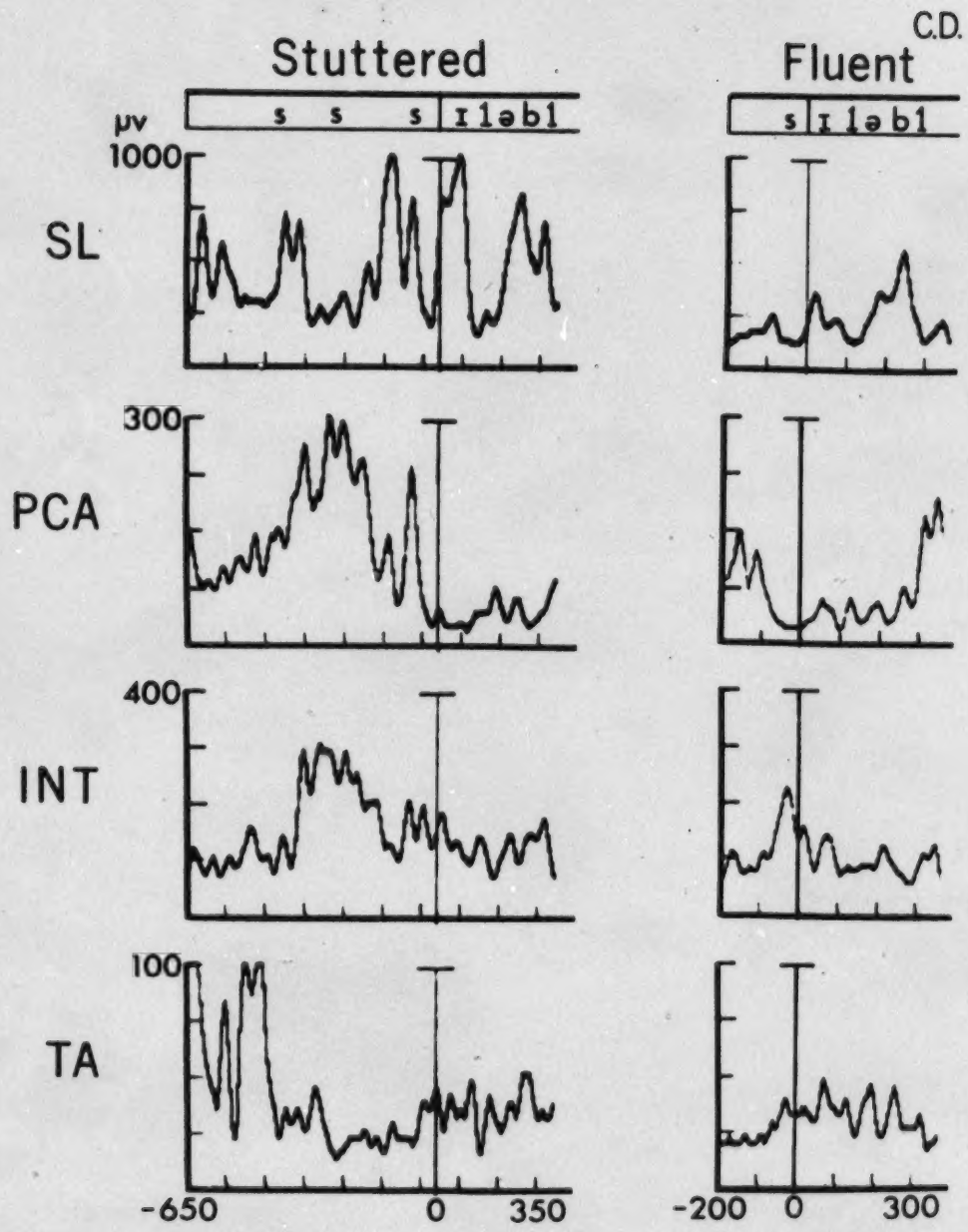


Figure 13: Comparison of muscle activity--superior longitudinal (SL), posterior cricoarytenoid (PCA), interarytenoid (INT), and thyroarytenoid (TA)--for subject C.D.'s stuttered and fluent utterances of the word "syllable."

230

The program (E$MGCORI) (Kewley-Port, 1975) used for these calculations plotted and correlated points at 5-msec intervals. Correlations were plotted for the time period between the first activity of the SL and PCA for the [s] and the onset of voicing for the vowel [ɪ]. Coefficients of correlation were calculated for 49 utterances of the words "syllable" and "syllables." As previously discussed, the experimenter had judged 23 of these utterances to be stuttered and 26 to be fluent.

Of the 23 utterances judged stuttered, 20 yielded positive correlations and 3 yielded negative correlations; while of the 26 utterances judged fluent, 19 yielded negative correlations and 7 yielded positive correlations. These findings are graphically illustrated in Figure 14.

In Figure 14, the 23 stuttered utterances are shown on the top half of the graph; while the 26 fluent utterances are shown on the lower half. All positive correlations are shown to the right of center, and negative correlations to the left. There is a significant positive correlation between abductor and adductor activity for the stuttered utterances ($p < .01$, sign test); there is a significant negative correlation between abductor and adductor activity for the fluent utterances ($p < .05$, sign test).

CONCLUSIONS

The results of the present study generate and support the following statements:

1. A laryngeal component of stuttering clearly exists.

2. Abnormal laryngeal muscle activity accompanied stuttering in all four subjects examined.

3. Two aspects of abnormal laryngeal muscle activity in stuttering are (a) high levels of muscle activity and (b) disruption of abductor-adductor reciprocity.

4. The cooccurrence of the three phenomena--(a) high levels of laryngeal muscle activity, (b) disrupted abductor-adductor reciprocity, and (c) perceived stuttering blocks--would support the hypothesis that the three are intimately related.

DISCUSSION

Generalization of Findirgs

The EMG results derived from four subjects take on additional significance when viewed in relation to the other physiological studies of laryngeal functioning in stuttering (Chevrie-Muller, 1963; Fujita, 1966; Ushijima et al., 1965; Conture, Brewer, and McCall, 1974). The picture emerging from these experiments (which were conducted independently, used a variety of instrumentations, and studied stutterers of three races who spoke three different languages) is consistent and supports the view that laryngeal involvement in stuttering is not an idiosyncratic phenomenon (Freeman, 1975).

FIGURE 14

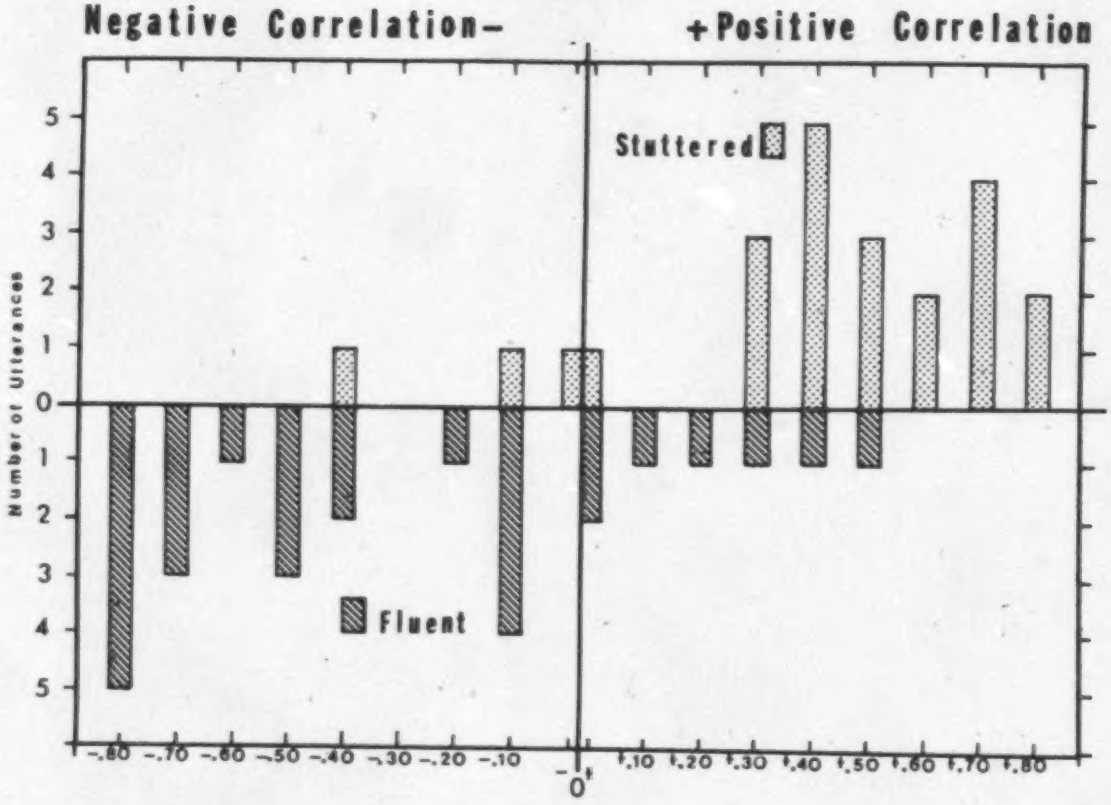


Figure 14: Comparison of abductor-adductor correlations for subject C.D.'s 26 fluent and 23 stuttered utterances of the first syllable of the words "syllable" and "syllables."

In most cases, the present study has verified hypotheses of researchers who used indirect approaches for studying phonation in stuttering. Adams and Reis (1971, 1974), Adams and Hayden (1974), Adams, Riemenschneider, Metz, and Conture (1974), and Agnello (1974), all predicted the initiation-of-phonation problem, demonstrated by subject C.D. and correlated with disrupted abductor-adductor reciprocity. If these investigators are correct in their interpretation, then they are observing indirectly in their subjects the same types of abnormal muscle activity studied directly in the present research.

Comments on Levels of Muscle Activity

The data relating to differences in levels of muscle activity may be interpreted in two ways, depending on the hypothesis espoused by the discussant. Both viewpoints are worthy of consideration.

The first hypothesis assumes that a moment of stuttering is accompanied by higher levels of muscle activity. It also assumes that the higher average levels found for passages in which stuttering moments occur are the result of averaging the high peak levels for the blocks with the normal base levels accompanying the nonstuttered speech. Certainly the results of the present research support the first contention of this hypothesis, namely, the stuttered utterance of a word is accompanied by levels of laryngeal muscle activity higher than those accompanying the fluent utterance of the same word (Figures 4, 5, 10, 11, 12, 13). However, if the raw EMG data (exemplified in Figures 1-3) is inspected closely, it becomes apparent that phrases within the stuttered readings in which no identifiable blocks occur are accompanied by levels of muscle activity that are higher than those accompanying the utterance of the same phrase in the fluent reading. Within the frame of this hypothesis, the higher levels accompanying the words on which there is no identifiable blocking can be explained in one of two ways: (1) by expanding time constraints on the moment of stuttering to include events that precede or follow the identifiable block, or (2) by assuming that in addition to the identified blocks, the stutterer is also experiencing a number of moments of stuttering, or minimal blocks, that are not recognized by the listener.

The second hypothesis assumes that the stutterer in specific communicative environments habitually attempts to phonate while maintaining higher than normal levels of laryngeal muscle activity. The high levels are viewed as being counterproductive in fluent utterance of sequential speech segments; and it is assumed that if the levels exceed some critical value, they will lead to a breakdown in fluency, that is, a moment of perceived stuttering. The data demonstrating lower levels of activity for the readings under the fluency-evoking conditions can be interpreted as supporting this hypothesis. The finding of higher levels for phrases that occur in the stuttered reading, but do not include identifiable blocks, would also support this line of reasoning.

Differentiation between the two hypotheses is difficult because both would predict similar patterns of correlation between levels of laryngeal muscle activity and occurrence of moments of stuttering. Both would predict that the highest levels of activity would coincide with identifiable blocks; both would predict increases in levels of activity during the time periods preceding identifiable blocks; and both would predict lower levels of activity during periods of fluency. The generally elevated baseline of activity during fluent utterance between blocks, which would be predicted by the second hypothesis, might be

testable if it were possible to define the temporal parameters of a given "moment of stuttering." However, if a "moment of stuttering" is viewed as including events that precede or follow the identifiable block by unspecified time periods, it becomes difficult or impossible to define the beginning or the end of a given "moment of stuttering." Although investigations of the temporal relationship between identifiable blocks and levels of laryngeal muscle activity are being conducted, no experimental method for testing the differential validity of these two hypotheses has yet been devised. On the other hand, it is also important to note that the two hypotheses are neither incompatible nor mutually exclusive.

Comments on Disrupted Reciprocity

As described by Sherrington (1909), "reciprocal inhibition" facilities coordinated movement by agonist muscles through relaxation of antagonist muscles. As demonstrated by Travill and Basmajian (1961), the antagonist in a muscle pair usually relaxes completely while the agonist is active. Studies of normal subjects, and indeed, recordings of the induced fluency readings of the stuttering subjects, show highly consistent reciprocity between the abductor (PCA) and the adductor group, particularly the INT. It is possible that whispered speech may be produced by simultaneous contraction of the PCA and some adductor muscles;[1] but for normal phonation the effects of abductor-adductor cocontraction are clearly counterproductive. From the data collected on D.M. and C.D., strong cocontraction of the laryngeal antagonists appears incompatible with normal phonation. In many instances, cocontraction occurred during a silent period just prior to an utterance. When cocontraction occurred during sound production, audible disruptions accompanied the event. For both subjects, the termination of cocontraction was almost invariably followed (50 to 150 msec) by a fluent sounding utterance.

Normal, fluent utterance of a CV syllable requires a specific change of laryngeal muscle tension pattern (this is true even if the consonant is voiced), and a specific change in glottal state (glottal constriction is different for consonants and vowels) within constrained time limits. Interpretation of the EMG evidence suggests that the effect of cocontraction was to prevent, delay, or inhibit the normal transition from the consonant into the vowel.

REFERENCES

Adams, M. R. and R. Hayden. (1974) Stutterers' and nonstutterers' ability to initiate and terminate phonation during nonspeech activities. Paper presented at the Annual Convention of the American Speech and Hearing Association, Las Vegas, Nev., 5-8 November.

Adams, M. R. and J. Hutchinson. (1974) The effects of three levels of auditory masking on selected vocal characteristics and the frequency of dysfluency of adult stutterers. J. Speech Hearing Res. 17, 682-688.

Adams, M. R. and R. Reis. (1971) The influence of the onset of phonation on the frequency of stuttering. J. Speech Hearing Res. 14, 639-644.

Adams, M. R. and R. Reis. (1974) The influence of the onset of phonation on the frequency of stuttering: A replication and re-evaluation. J. Speech Hearing Res. 17, 752-754.

[1] Thomas Shipp, 1975: personal communication.

Adams, M. R., S. Riemenschneider, D. E. Metz, and E. G. Conture. (1974) Voice onset and articulatory constriction requirements in a speech segment, and their relation to the amount of stuttering adaptation. Paper presented at the Annual Convention of the American Speech and Hearing Association, Las Vegas, Nev., 5-8 November.

Agnello, J. (1971) Transitional features of stutterers and nonstutterers. ASHA: Journal of the American Speech and Hearing Association 12(A).

Agnello, J. (1974) Laryngeal and articulatory dynamics of dysfluency interpreted within a vocal tract model. In Vocal Tract Dynamics and Dysfluency, ed. by L. M. Webster and L. C. Furst. (New York: Speech and Hearing Institute).

Arnott, G. Niel. (1828) Elements of Physics as reported in Hunt (1861).

Basmajian, J. V. and G. A. Stecko. (1962) A new bipolar indwelling electrode for electromyography. J. Appl. Physiol. 17, 849.

Bigland, B. and O. C. J. Lippold. (1954) The relation between force, velocity and integrated electrical activity in human muscles. J. Physiol. 123, 214-224.

Brenner, N. C., W. H. Perkins, and G. A. Soderberg. (1972) The effect of rehearsal on frequency of stuttering. J. Speech Hearing Res. 15, 474-482.

Chevrie-Muller, C. (1963) A study of laryngeal function in stutterers by the glottal-graphic method. In Proc. VII Congress de la Societé Française de Medicine de la Voix et de la Parole, Paris.

Conture, E. G. (1974) Some effects of noise on the speaking behavior of stutterers. J. Speech Hearing Res. 17, 714-723.

Conture, E. G., D. W. Brewer, and G. N. McCall. (1974) Laryngeal activity during the moment of stuttering: Some preliminary observations. Paper presented at the Annual Convention of the American Speech and Hearing Association, Las Vegas, Nev., 5-8 November.

Faaborg-Anderson, K. (1957) Electromyographic investigation of intrinsic laryngeal muscles in humans. Acta Physiol. Scand., Suppl. 41, 140.

Freeman, F. J. (1975) The stuttering larynx: An electromyographic study of laryngeal muscle activity accompanying stuttering. Unpublished doctoral dissertation, City University of New York.

Freeman, F. J., M. F. Dorman, T. Ushijima, and S. Niimi. (1975) Laryngeal dysfunction in stuttering: EMG and fiberoptic studies. Unpublished manuscript.

Fujita, K. (1966) Pathophysiology of the larynx from the viewpoint of phonation. J. Japan. Soc. Otorhinolaryngol. 69, 459.

Gay, T. and H. Hirose. (1973) Effect of speaking rate on labial consonant production: A combined electromyographic/high-speed motion picture study. Phonetica 27, 44-56.

Gay, T., M. Strome, H. Hirose, and M. Sawashima. (1972) Electromyography of the intrinsic laryngeal muscles during phonation. Ann. Otol. Rhinol. Laryngol. 81, 401-408.

Hirano, M. and J. Ohala. (1969) Use of hooked-wire electrodes for electromyography of the intrinsic laryngeal muscles. J. Speech Hearing Res. 12, 362-373.

Hirano, M., J. Ohala, and W. Vennard. (1970) Regulation of register, pitch and intensity of voice. Folia Phoniat. 22, 1-20.

Hirose, H. (1971) Electromyography of the articulatory muscles: Current instrumentation and technique. Haskins Laboratories Status Report on Speech Research SR-25/26, 73-86.

Hirose, H. (1974) Functional differentiation of the glottal adductors. Japan. J. Otol. 77, 46-57.

Hirose, H. and T. Gay. (1972) The activity of the intrinsic laryngeal muscles in voicing control: An electromyographic study. Phonetica 25, 140-164.
Hirose, H. and T. Gay. (1973) Laryngeal control in vocal attack: An electromyographic study. Folia Phoniat. 25, 203-213.
Hirose, H. and T. Ushijima. (1974) The function of the posterior cricoarytenoid in speech articulation. Haskins Laboratories Status Report on Speech Research SR-37/38, 99-107.
Hunt, J. (1861) Stammering and Stuttering: Their Nature and Treatment, 1967 ed. (London: Hafner Publishing Co.).
Kenyon, E. L. (1943) The etiology of stammering: The psychophysiologic facts which concern the production of speech sounds and of stammering. J. Speech Hearing Dis. 8, 337-348.
Kewley-Port, D. (1973) Computer processing of EMG signals at Haskins Laboratories. Haskins Laboratories Status Report on Speech Research SR-33, 173-184.
Kewley-Port, D. (1974) An experimental evaluation of the EMG data processing system: Time constant choice for digital integration. Haskins Laboratories Status Report on Speech Research SR-37/38, 65-72.
Kuehn, D. P. (1973) A cinefluorographic investigation of articulatory velocities. Unpublished doctoral thesis, University of Iowa.
Moravek, M. and J. Langova. (1967) Problems of the development of the initial tonus in stuttering. Folia Phoniat. 19, 109-116.
Müller, J. P. (1833) Elements of Physiology, trans. by Baly (1857), reported in Hunt.
Port, D. K. (1971) The EMG data system. Haskins Laboratories Status Report on Speech Research SR-25/26, 67-72.
Schwartz, M. (1974) The core of the stuttering block. J. Speech Hearing Dis. 39, 169-177.
Sherrington, C. S. (1909) Reciprocal innervation of antagonistic muscles. Fourteenth note. On double reciprocal innervation. Proc. Royal Soc. B81, 249-268.
Shipp, T. and R. McGlone. (1971) Laryngeal dynamics associated with voice frequency change. J. Speech Hearing Res. 4, 761-768.
Stromstra, C. (1965) A spectrographic study of dysfluencies labeled as stuttering by parents. De Therapia Vocis et Loquellae 1, 317-320.
Travill, A. and J. V. Basmajian. (1961) Electromyography of the supinators of the forearm. Anat. Rec. 139, 557-560.
Ushijima, T., G. Kamiyama, H. Hirose, and S. Niimi. (1965) Articulatory movements of the larynx during stuttering (a film produced at the Research Institute of Logopedics and Phoniatrics, Faculty of Medicine, University of Tokyo).
Wingate, M. E. (1969) Sound pattern in "artificial" fluency. J. Speech Hearing Res. 12, 677-686.
Wingate, M. E. (1970) Effect on stuttering of changes in audition. J. Speech Hearing Res. 13, 861-873.
Wyke, B. (1971) The neurology of stammering. J. Psychosomatic Res. 15, 423-432.

II. PUBLICATIONS AND REPORTS

Abramson, A. S. (1976) Static and dynamic acoustic cues in distinctive tones. Journal of the Acoustical Society of America, Suppl. 59, S42(A).

Blechner, M. J., R. S. Day, and J. E. Cutting. (1976) Processing two dimensions of nonspeech stimuli: The auditory-phonetic distinction reconsidered. Journal of Experimental Psychology: Human Perception and Performance 2, 257-266.

Cutting, J. E. (1976) Auditory and linguistic processes in speech perception: Inferences from six fusions in dichotic listening. Psychological Review 83, 114-140.

Healy, A. F. (1976) Detection errors on the word the: Evidence for reading units larger than letters. Journal of Experimental Psychology: Human Perception and Performance 2, 235-242.

Healy, A. F. and J. E. Cutting. (1976) Units of speech perception: Phoneme and syllable. Journal of Verbal Learning and Verbal Behavior 15, 73-83.

Mermelstein, P. (1976) The syntax of acoustic segments. Conference Record, 1976 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 33-36.

Mermelstein, P. and S. Levinson. (1976) Speech recognition: Acoustic, phonetic and formal-language models. In Proceedings of the Fourth New England Bioengineering Conference, ed. by S. Saha. (New York: Pergamon Press), pp. 475-477.

III. APPENDIX

DDC (Defense Documentation Center) and ERIC (Educational Resources Information Center) numbers:

SR-21/22 to SR-44

Status Report		DDC	ERIC
SR-21/22	January - June 1970	AD 719382	ED-044-679
SR-23	July - September 1970	AD 723586	ED-052-654
SR-24	October - December 1970	AD 727616	ED-052-653
SR-25/26	January - June 1971	AD 730013	ED-056-560
SR-27	July - September 1971	AD 739339	ED-071-533
SR-28	October- December 1971	AD 742140	ED-061-837
SR-29/30	January - June 1972	AD 750001	ED-071-484
SR-31/32	July - December 1972	AD 757954	ED-077-285
SR-33	January - March 1973	AD 762373	ED-081-263
SR-34	April - June 1973	AD 766178	ED-081-295
SR-35/36	July - December 1973	AD 774799	ED-094-444
SR-37/38	January - June 1974	AD 783548	ED-094-445
SR-39/40	July - December 1974	AD A007342	ED-102-633
SR-41	January - March 1975	AD A103325	ED-109-722
SR-42/43	April - September 1975	AD A018369	ED-117-770
SR-44	October - December 1975	AD A023059	ED-119-273

AD numbers may be ordered from: U.S. Department of Commerce
National Technical Information Service
5285 Port Royal Road
Springfield, Virginia 22151

ED numbers may be ordered from: ERIC Document Reproduction Service
Computer Microfilm International Corp. (CMIC)
P.O. Box 190
Arlington, Virginia 22210

Haskins Laboratories Status Report on Speech Research is abstracted in Language and Behavior Abstracts, P.O. Box 22206, San Diego, California 92122.

UNCLASSIFIED

Security Classification

DOCUMENT CONTROL DATA - R & D

Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified

1. ORIGINATING ACTIVITY (Corporate author)	2a. REPORT SECURITY CLASSIFICATION
Haskins Laboratories, Inc. 270 Crown Street New Haven, Connecticut 06510	Unclassified 2b. GROUP N/A

3. REPORT TITLE

Haskins Laboratories Status Report on Speech Research, No. 45/46, January - June 1976

4. DESCRIPTIVE NOTES (Type of report and, inclusive dates)

Interim Scientific Report

5. AUTHOR(S) (First name, middle initial, last name)

Staff of Haskins Laboratories; Alvin M. Liberman, P.I.

6. REPORT DATE	7a. TOTAL NO. OF PAGES	7b. NO. OF REFS
May 1976	241	316

8. CONTRACT OR GRANT NO	9a. ORIGINATOR'S REPORT NUMBER(S)
DE-01774 HD-01994 V101(134)P-342 N00014-76-C-0591 DAAB03-75-C-0419(L433) N01-HD-1-2420 RR-5596	SR-45/46 (1976)
	9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report) None

10. DISTRIBUTION STATEMENT

Distribution of this document is unlimited.*

11. SUPPLEMENTARY NOTES	12. SPONSORING MILITARY ACTIVITY
N/A	See No. 8

13. ABSTRACT

This report (1 January - 30 June 1976) is one of a regular series on the status and progress of studies on the nature of speech, instrumentation of its investigation, and practical applications. Manuscripts cover the following topics:

Exploring Relations between Reading and Speech
Interpreting Error Pattern in Beginning Reading
Comments on Session: Perception and Production of Speech II; Conference on Origins and Evolution of Language and Speech
Consonant Environment Specifies Vowel Identity
What Information Enables Listener to Map Talker's Vowel Space?
Identification Dichotic Fusions
Discrimination Dichotic Fusions
Coperception: Two Further Preliminary Studies
"Posner's Paradigm" and Categorical Perception: Negative Study
Weak Syllables in Primitive Reading-Machine Algorithm
Control Fundamental Frequency, Intensity, Register of Phonation
Effect of Delayed Auditory Feedback on Phonation: Electromyographic Study
Some Aspects of Coarticulation
Function of Strap Muscles in Speech
Laryngeal Muscle Activity in Stuttering

DD FORM 1 NOV 65 1473 (PAGE 1)

S/N 0101-807-6811

*This document contains no information not freely available to the general public. It is distributed primarily for library use.

UNCLASSIFIED

Security Classification

A-31408

14 KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Reading and Speech - Relations						
Reading Errors - Interpretation						
Speech Production: Language Evolution						
Vowel Identity: Consonant Environment						
Vowel Space - Mapping						
Dichotic Fuisons - Identification						
Dichotic Fusions - Discrimination						
Coperception						
Categorical Perception - Posner's Paradigm						
Syllables Weak - Algorithm						
Phonation, Register, Fundamental Frequency, Intensity						
Phonation, Delayed Feedback - Study						
Coarticulation - Aspects						
Strap Muscles - Function						
Stuttering - Muscle Activity						